dortmund university

fakultät statistik

Statistical methods for complex data structures: change detection and goodness-of-fit testing for stochastic networks and an online monitoring framework for text data

Dissertation

by

Jonathan Flossdorf

in partial fulfillment of
the requirements for the degree of
Doktor der Naturwissenschaften (Dr. rer. nat.)

Submitted: Dortmund, August 2024
Primary referee: Prof. Dr. Carsten Jentsch
Secondary referee: Jun.-Prof. Dr. Alexander Kreiss
Third referee: Prof. Dr. Roland Fried

Dissertation
in partial fulfillment of
the requirements for the degree of
Doktor der Naturwissenschaften (Dr. rer. nat.)

Submitted to the
Department of Statistics of the
TU Dortmund University

Dortmund, August 2024

Primary referee: Prof. Dr. Carsten Jentsch
Secondary referee: Jun.-Prof. Dr. Alexander Kreiss
Third referee: Prof. Dr. Roland Fried

Acknowledgments

I would like to deeply thank Carsten Jentsch for giving me the opportunity to work in his group and for introducing me to the world of science. I highly appreciate his great confidence in my work and his helpful advice during all this time. I have learnt a lot and I could not have wished for more!

I would also like to thank all my co-authors and colleagues for their useful insights and the humorous discussions we have had over the past few years.

Lastly, this work is devoted to my wife, whose endless support is the sole reason why I have been able to complete this project and why I may enjoy the most wonderful life.

Contents

Abstract

In modern times, not only the amount of data is growing, but also its complexity leading to more sophisticated types of data. A popular example is the conception of given structures as networks. Because network data, even in their simplest form, contain considerably more information than traditional data, they are methodologically challenging to handle and analyse - even more so as sample sizes increase or covariates are added.

This cumulative dissertation addresses the challenges of descriptive and inferential statistics in the context of complex and high-dimensional structures with a main focus on stochastic networks. A special centre of attention lies in maintaining flexibility in practical application, i.e. imposing as few restrictions on the network structure as possible. As a first step, the peculiarities of network data and the question of how to characterise differences between given networks are considered, leading to a categorisation of potential change types in network structure. Further, the suitability of different network metrics, which characterise a network by a scalar value, is investigated in a comprehensive framework. These findings are then embedded in a change point detection scheme for dynamic networks. In a second work, these findings are extended to a multivariate setup where different metrics are used simultaneously to extract as much information as possible from the dynamic network. In this context, the interplay between multivariate metric sets and different types of parametric and non-parametric control charts is extensively discussed. Third, a novel class of goodness-of-fit tests is introduced, which includes a wide range of test statistics and allows to decide whether an underlying network is generated by a homogeneous Erdös-Rényi model, which is typically used as a benchmark model due to its simplicity and handiness, or by a more sophisticated alternative.

In the latter stages of this thesis, attention is turned to the field of textual data, which, due to its complexity, requires overcoming similar challenges. A change detection scheme is developed that allows to automatically monitor the evolution of topics identified from a dynamic topic modeling approach. In order to increase the flexibility and applicability of this procedure, it is further embedded in a rich visualisation scheme that enables advanced interpretation possibilities and deeper analysis options for the user.

Zusammenfassung

In der heutigen Zeit wächst nicht nur die Menge an Daten, sondern auch ihre Komplexität, was zu immer komplexeren Datentypen führt. Ein beliebtes Beispiel ist die Konzeption von gegebenen Strukturen als Netzwerke. Da Netzwerkdaten, selbst in ihrer einfachsten Form, wesentlich mehr Informationen enthalten als herkömmliche Daten, sind sie methodisch anspruchsvoll zu handhaben und zu analysieren - umso mehr, je größer die Stichproben sind oder je mehr Kovariablen hinzugefügt werden.

Diese kumulative Dissertation befasst sich mit den Herausforderungen deskriptiver und inferenzstatistischer Methoden für komplexe und hochdimensionale Datenstrukturen mit einem Schwerpunkt auf stochastischen Netzwerken. Ein besonderes Augenmerk liegt dabei auf der Wahrung von Flexibilität in der praktischen Anwendung, d.h. möglichst wenig Annahmen an die Netzwerkstruktur zu stellen. In einem ersten Schritt werden die Besonderheiten von Netzwerkdaten und die Frage, wie Unterschiede zwischen gegebenen Netzwerken charakterisiert werden können, betrachtet, was zu einer Kategorisierung von möglichen Veränderungen (‘changes”) in der Netzwerkstruktur führt. Weiterhin wird die Eignung verschiedener Netzwerkmetriken, die ein Netzwerk durch einen skalaren Wert charakterisieren, in Bezug auf Erkennung solcher Veränderungen in einem umfassenden Rahmen untersucht. Diese Erkenntnisse werden anschließend in ein Verfahren zur Strukturbrucherkennung in dynamischen Netzwerken eingebettet. In einer zweiten Arbeit werden diese Erkenntnisse zu einem multivariaten Ansatz ausgeweitet, bei dem verschiedene Metriken gleichzeitig verwendet werden, um so viele Informationen wie möglich aus dem dynamischen Netzwerk zu extrahieren. In diesem Zusammenhang wird das Zusammenspiel zwischen multivariaten Metriksätzen und verschiedenen Arten von parametrischen und nicht-parametrischen Kontrollkarten ausführlich diskutiert. Als drittes wird eine neue Klasse von Anpassungstests eingeführt, die eine breite Palette von Teststatistiken umfasst. Diese Tests ermöglichen es zu entscheiden, ob ein zugrunde liegendes Netzwerk durch ein einfaches homogenes Erdös-Rényi-Modell modelliert werden kann, welches aufgrund seiner Handlichkeit üblicherweise als Referenzmodell dient, oder ob eine anspruchsvollere, heterogene Alternative eine signifikant bessere Anpassungsgüte erreicht.

In den letzten Abschnitten dieser Arbeit wird die Aufmerksamkeit auf den Bereich der Textdaten gelenkt, der aufgrund seiner Komplexität die Bewältigung ähnlicher Herausforderungen erfordert. Es wird ein online monitoring Verfahren entwickelt, das es ermöglicht, die Entwicklung von Themen, die mit Hilfe eines dynamischen Themenmodellierungsansatzes identifiziert wurden, automatisch zu überwachen. Um die Flexibilität und Anwendbarkeit dieses Verfahrens zu erhöhen, wird es in ein umfangreiches Visualisierungsschema eingebettet, das dem Benutzer fortgeschrittene Interpretationsmöglichkeiten und tiefere Analyseoptionen bietet.

List of publications

The following peer-reviewed and published papers are part of this cumulative dissertation. They have been reused with the permission of the copyright holder.

Chapter 2: Flossdorf, Jonathan & Carsten Jentsch (2021): "Change detection in dynamic networks using network characteristics", In: *IEEE Transactions on Signal and Information Processing over Networks* 7, pp. 451–464, DOI: 10.1109/TSIPN.2021.3094900. ©2021 IEEE. Reprinted, with permission.

Chapter 3: Flossdorf, Jonathan, Roland Fried & Carsten Jentsch (2023): "Online monitoring of dynamic networks using flexible multivariate control charts", In: *Social Network Analysis and Mining* 13, Article No. 87, DOI:10.1007/s13278-023-01091-y.

Chapter 4: Brune, Barbara, Jonathan Flossdorf & Carsten Jentsch (2024): "Goodness-of-fit testing based on graph functionals for homogeneous Erdös-Rényi graphs", *Scandinavian Journal of Statistics*, 1–49, DOI:10.1111/sjos.12750.

Chapter 5: Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf & Carsten Jentsch (2022): "Dynamic change detection in topics based on rolling LDAs", In: *Proceedings in the Text2Story'22 Workshop*. Vol. 3117.CEUR-WS, pp. 5–13.

Chapter 6: Krause, Cedric, Jonas Rieger, Jonathan Flossdorf, Carsten Jentsch, and Fabian Beck (2023). "Visually Analyzing Topic Change Points in Temporal Text Collections". In: *Vision, Modeling, and Visualization*. Ed. by Michael Guthe and Thorsten Grosch. The Eurographics Association. ISBN: 978-3-03868-232-5. DOI: 10.2312/vmv.20231231.

Abbreviations

This list includes all abbreviations used in this thesis and some important, recurring notations.

$\emptyset$	null graph
$\|.\|$	matrix norm
α	significance level (false alarm rate in control chart context)
A	adjacency matrix
$\|A\|_F^2$ and F	Frobenius norm of A
AIC	Akaike Information Criterion
ARIMA	autoregressive integrated moving average
ARL_0	in-control average run length
ARL_1	post-change average run length (average detection length)
aut(.)	number of automorphisms
AvP	average path length
BIC	Bayesian Information Criterion
C_3	triangle (complete graph of order 3)
CUSUM	cumulative sum
$c(v)$	centrality score of a node v
cos(.)	cosine similarity
$\mathcal{D}$	dynamic network
$\mathbb{E}(.)$	expected value
E	edge set
E_n	eigenvector centrality
ER	Erdös-Rényi
EWMA	exponential weighted moving average
G	graph
$\mathcal{G}(n,p)$ and $\mathcal{G}_{\text{ER}}(n)$	(homogenous) Erdös-Rényi graph
$\mathcal{G}(n,\boldsymbol{P})$ and $\mathcal{G}_{\text{HER}}(n)$	heterogenous Erdös-Rényi graph
GHRG	Generalized Hierarchical Random Graph

GLC	global link change
GNC	global node change
HER	heterogenous Erdös-Rényi
LDA	Latent Dirichlet Allocation
LLC	local link change
LNC	local node change
m	number of links
MEWMA	multivariate exponential weighted moving average
MTC	mixed-type change
n	number of nodes
NLP	Natural language processing
$\boldsymbol{P}$	matrix of link probabilities
p	(link) probability
p_{mean}	mean probability
P_2	graph with two nodes and one edge
P_3	two-star (graph with three nodes and two edges)
S	Spectral norm
$\boldsymbol{S}$	covariance matrix
$S_n(.)$	centered subgraph count
SBM	stochastic block model
SPC	statistical process control
$T_n(.)$	(raw) subgraph count
Tr	transitivity (clustering coefficient)
V	vertex set
$\mathbb{V}\text{ar}(.)$	variance
V_n	degree variance
X_n	graph functional
$(x)_y$	descending factorials

1

Introduction

As a result of the growing volume and complexity of data, the conception of given structures as networks is becoming increasingly popular. One reason for this is its broad applicability and flexibility, as network analysis enables the representation, understanding, modelling and interpretation of relationships between any type of entity in almost any field of study. Typical examples that underline this diversity are trade flows between different countries, friendships on social media platforms, the transmission of infectious diseases, or the tracking of orders in supply chains.

According to Newman, 2018, we can loosely categorise four different application areas of networks. The first category comprise technological networks, which includes the physical networks that underpin a modern technological society, with the Internet being the most popular example. It also includes all types of transport networks, with delivery and distribution processes as special cases. Secondly, there are information networks, which consist of pieces of data that are linked together in some way. These include hyperlinks in the World Wide Web or citation networks in research communities. Thirdly, there are social networks, which are arguably the most widely studied. The entities in these networks are always people, resulting, e.g., in friendship networks or communication networks. Lastly, there are biological networks, which represent systems as complex sets of binary interactions or relationships between different biological entities, such as protein-protein interactions or metabolic networks.

1.1 Network data

Generally speaking, network data is extremely diverse. It can be enriched with almost any kind of information and is able to map complex structures and address functional relationships. Even in its simplest form, it is very different from traditional statistical data setups and becomes even more complex as additional information is used. It is therefore important to introduce standardised terminology and to gain an understanding of the basic concepts. Only then can more in-depth statistical analysis be carried out.

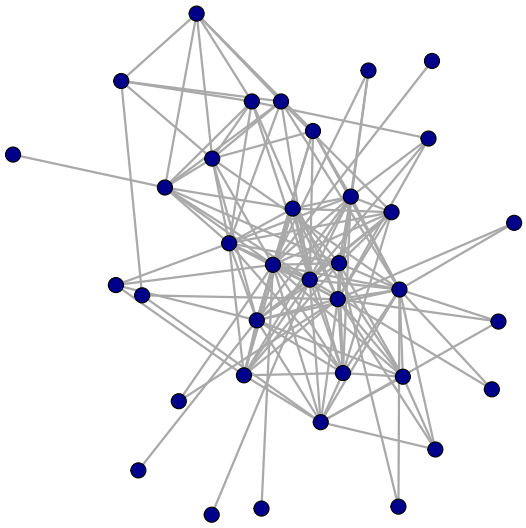


Figure 1: Example for a simple graph using the hospital data.

1.1.1 Simple graphs

All networks consist of vertices (also called nodes) and edges (also called links). The vertices represent the entities in the investigated system and the edges characterize the relations between the vertices. Ultimately, a graph G can be defined by its vertex set $V = \{v_1, ..., v_n\}$ and its edge set $E = \{e_1, ..., e_m\}$. We first consider the most parsimonious graph setup, where G consists only of V and E, and an undirected edge is drawn between two vertices if there is a connection between them. This setup is also called *simple graph*. An alternative representation of G which allows a more mathematically precise treatment is the adjacency matrix $A \in \mathbb{R}^{n \times n}$. For now, the entries of A take values of $a_{ij} = 1$ if there is a connection between vertices v_i and v_j, and $a_{ij} = 0$ otherwise with $i, j = 1, ..., n$. In a simple graph, the matrix A is symmetric, i.e. $a_{ij} = a_{ji}$, since the relation between two vertices is not directed, and the diagonal entries are 0 by construction, i.e. $a_{ii} = 0$, since it is assumed that there is no connection between a vertex and itself. To illustrate the concept of a simple graph, consider an example from social network science [Vanhems et al., 2013]: It includes records of contacts between patients and different types of health care workers in the geriatric ward of a hospital in the French city of Lyon in 2010. In this study, people in the hospital consented to wear RFID sensors on small identification badges, which made it possible to record when any two of them were in face-to-face contact (i.e. within 1-1.5 m of each other) during a 20-second time interval. The simple graph visualised in Figure 1 shows the resulting network for the time period of

Monday 6 December from 1pm to 3pm. It shows 21 unlabelled vertices that are connected by 54 edges. On a structural level, there clearly are vertices that are important for the network as they share many connections with other ones, while there are also some that only have a few connections - some even have a single one only. This observation leads to the class of centrality metrics, which provide a way to measure parts of the structure of a given network with a scalar value in order to make them mathematically better comparable. Centrality metrics are calculated vertex-wise such that each vertex in the network has an individual score. These scores measure the centrality, i.e. the importance, of a vertex in the whole network. As there exist different concepts of centrality, there are a variety of possible metrics. Arguably the most simple one is the degree centrality d_i that just counts the number of connections a vertex v_i is part of. More sophisticated concepts also take into account the importance of neighbouring nodes, so that relationships to high scoring nodes are more valuable than those to low scoring ones. A full discussion of these metrics and their suitability for different statistical analysis tasks is given later in the main body of this paper. For now, we can see from Figure 1 that the degree centrality between nodes varies considerably, resulting in a rather centralised layout with a few hub nodes.

1.1.2 Covariates on vertices

An advantage of network data is that the user can attach nearly any additional information to it. One example is to characterise the entities in the system, i.e. to attach covariates to the nodes. In the simplest case, this could be characterising the nodes by an ID number to make them more tractable. Other typical covariates include gender or age when the entities are people, coordinates or countries when the entities are locations, as well as sizes or types of cells in biological applications. In the hospital data example, we have a given covariate called status, which specifies the role of the person in the hospital and can take the values {PAT: patient, NUR: nurse, MED: medical doctor, ADM: administrative staff}. Figure 2 visualises the resulting network with different vertex colours for the different levels of the applied covariate. We can now make sharper observations compared to the simple graph before. The visualisation shows that nurses are the most central nodes in the network, while patients seem to have less influence. This is also confirmed by the values of the degree centralities, aggregated over all 4 groups we have $d_{\text{NUR}} = 15$, $d_{\text{PAT}} = 2$, $d_{\text{MED}} = 9$ and $d_{\text{ADM}} = 9$. This behaviour is intuitively plausible. Nurses have contact with patients, but also with doctors and administration. They are therefore in contact with many people. A patient, on the other hand, is likely to have contact with only a small subset of some nurses and doctors - and perhaps other patients with whom they share a room. As we can see, using this additional information greatly increases flexibility and interpretation. However, it also makes the network more complex and difficult to

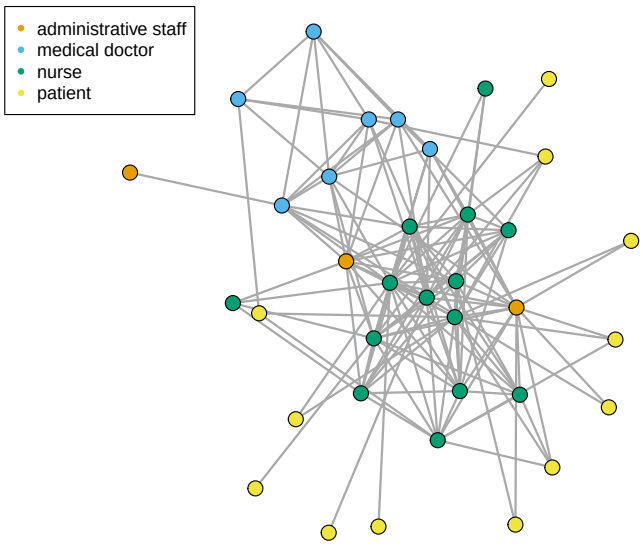


Figure 2: Expansion of the hospital data with a covariate on vertices.

manage from a statistical point of view. For example, we have to decide whether this additional information is important enough to include in the modelling.

1.1.3 Covariates on edges

As well as placing covariates at the vertices, we can also place them at the edges. This is typically done to characterise the intensity of a relationship. This could be the distance between locations or the number of goods transported in a value chain. Such concepts are also referred to as weighted edges, resulting in a *weighted graph*. In the hospital example, we can examine not only whether two people have been in contact, but also how often this has happened in the given time period. However, even this simple change might raise visualisation issues. Placing numerical values on the edges would make the visualisation confusing and hard to identify in a meaningful way. A simple alternative is to plot the thickness of the edges with their respective intensities, as shown in Figure 3. This allows us to draw further conclusions, e.g. identifying the patients who need more care or verifying that the communication between the doctors is quite intensive. The adjacency matrix A changes compared to the simple graph setup in the sense that non-binary values and even decimals are allowed. The entries a_{ij} then indicate the weights of the corresponding edges.

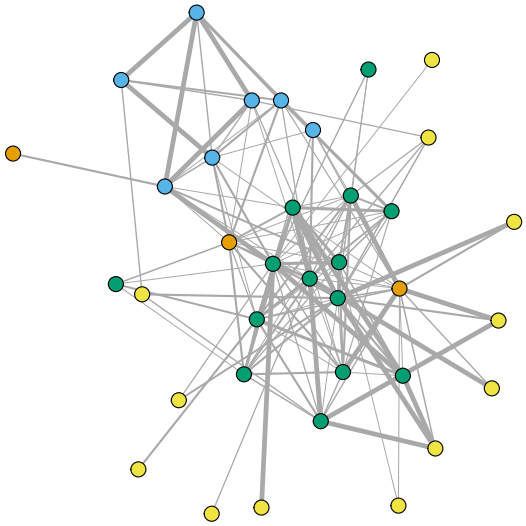
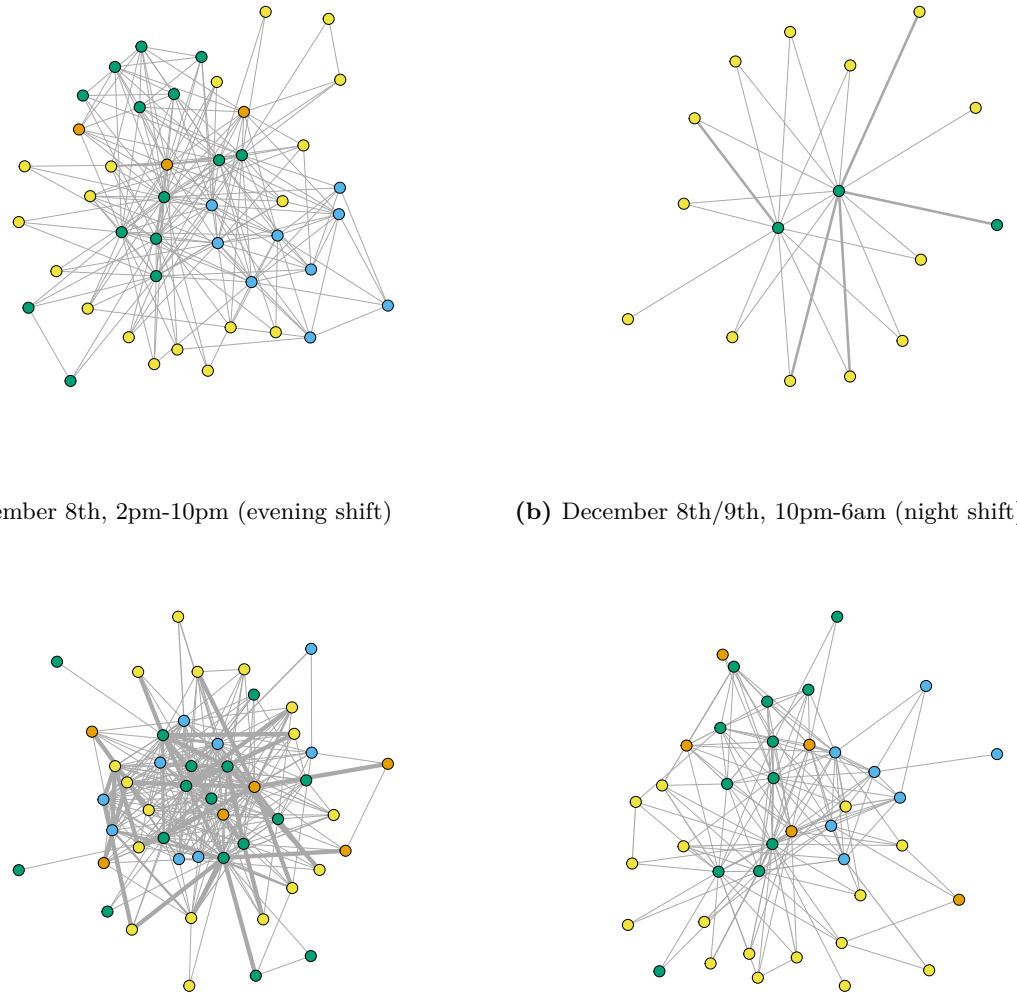


Figure 3: Expansion of the hospital data with a covariate on edges.

1.1.4 Dynamic Networks

Another extension is to observe a given network over time, which is then called a *dynamic network* $\mathcal{D}$. There are basically two ways of doing this. Firstly, we can attach a covariate to the edges, specifying the time at which the edge exists (temporal edges). This has the advantage of capturing all changes as we get a somewhat continuous representation of the network. However, this is mathematically difficult to handle, especially for high dynamics. A more convenient setup is to take snapshots of the network at different time points, so that the dynamic network consists of a sequence of graphs over time, i.e. $\mathcal{D} = \{G_1, ..., G_T\}$. This provides a more intuitive and better addressable setup. However, this is a discrete approach and it is crucial to choose the time points carefully so as to lose as little important information as possible. For the example of the hospital data, it seems natural to choose the time intervals according to the shifts of the medical staff. The result is shown in Figure 4. The most obvious observation is the greatly reduced amount of contact during the night shift. There is only some contact between patients and a few nurses and in-between nurses. There are no administrative staff in the network, as they do not work at night, and there are no medical doctors, as they are probably only on emergency duty during the night shift. The two evening shifts shown in Figure 4a and Figure 4d appear relatively similar at first glance. The early shift is the busiest as most medical treatment and doctor visits take place.



(a) December 8th, 2pm-10pm (evening shift)

(b) December 8th/9th, 10pm-6am (night shift)

(c) December 9th, 6am-2pm (early shift)

(d) December 9th, 2pm-10pm (evening shift)

Figure 4: Dynamic network visualization of the hospital data.

1.1.5 Further extensions

In addition to these possibilities, there are a number of other concepts that can extend statistical networks. A common example is the consideration of a *directed graph*, where edges can point from one vertex to another and vice versa, and are typically represented by an arrow. As a result, the edge set E can be twice as large as in an undirected graph, and the adjacency matrix A may no longer be symmetric. An example is a citation network, which represents the citations within a collection of scientific publications or authors. In this context, a directed edge is drawn when one publication or author cites another. Related to directed graphs is the concept of *self-loops*, which removes the restriction that $a_{ii} = 0$, thus allowing for the existence

of an edge between a vertex and itself (e.g. in an author-based citation network, when an author cites another work by himself). Another extension is offered by so-called *multigraphs*, where more than one relative component is considered, resulting in a graph in which there are different types of edges between the vertices of a given vertex set. Multigraphs are often used in the analysis of social networks, where it may be interesting to examine the relationship between two people at different levels, e.g. friendship and membership of the same sports club. Furthermore, a *bipartite graph* is a graph whose vertices can be divided into two disjoint and independent sets V_1 and V_2, i.e. each edge connects a vertex in V_1 to one in V_2. Such graphs often arise naturally when analysing relationships between two different classes of entities, e.g. athletes and sports clubs.

1.2 Outline and research questions

As these examples suggest, it is almost impossible to treat every single piece of information provided by a network. This is the case with any statistical task involving network data. In a descriptive analysis, it is hardly feasible to visualise all network information in a meaningful way while maintaining the clarity and expressiveness of the visualisation. Furthermore, describing the network by one or more characteristics will obviously lead to a loss of information at some point, since it is impossible to measure all system processes with a scalar-valued metric. In terms of inferential statistical analysis, network models must find the difficult balance of reliably representing complex processes in a high-dimensional system while avoiding overfitting.

1.2.1 Network metrics and change detection

A crucial aspect of any statistical approach to network analysis is therefore to identify and extract the useful information from the given network structure to answer the research questions, while neglecting any information that is not related to these questions. In this context, describing the network by a network metric plays an important role as it is the first step for almost any statistical approach. Obviously, for descriptive statistics, by characterising the network with a measurable function, but also for inferential statistics, as test statistics and therefore the performance of a test procedure, strongly depend on the choice of a network metric. The next two chapters of this thesis are devoted to investigating the suitability of typical network metrics for different types of network structures and statistical tasks. While Chapter 2 focuses on the univariate performance of these metrics, Chapter 3 discusses their multivariate combination in order to extract as much useful information as possible. Both analyses are embedded in a change point detection scheme for a dynamic network setup, considering both online and offline applications. To further increase the flexibility of the results, both parametric and

non-parametric frameworks are analysed. Regarding the hospital data example, the use of change point detection could be useful to answer the following questions

- Is there a lack of staff (medical doctors, nurses, administrative workers,...)?
- Is there a smooth transition between the shifts?
- Can the working proccesses be optimized?
- Is the emergency occupation during night shifts satisfactory?

Furthermore, the results can be used to find an appropriate test statistic also for non-sequential tests such as goodness-of-fit approaches or two-sample tests.

1.2.2 Goodness-of-fit testing

In light of these results, Chapter 4 provides a specific goodness-of-fit framework that decides whether an observed network is generated by a particular class of models. Specifically, we focus on so-called Erdös-Rényi models.

> Let $G = (V, E)$ be a random graph on n vertices with adjacency matrix A and let $p \in [0, 1]$ be the connection probability between two vertices. Then, we call G an *Erdös-Rényi graph* $\mathcal{G}(n, p)$, if the edges are realizations of stochastically independent and identically Bernoulli distributed random variables. That is, for $1 \leq i < j \leq n$, we have $A_{ij} \sim \text{Bin}(1, p)$ with $A_{ji} := A_{ij}$, and $A_{ii} := 0$ for all i.

This is the most popular and arguably the most parsimonious model, being characterised by only two parameters, n and p. In practice, it is often used as a benchmark model, meaning that a more sophisticated model should achieve a significantly better fit in order to be an option for modelling the underlying data. The Erdös-Rényi model can be generalised to more heterogeneous variants.

> Let $G = (V, E)$ be a random graph on n vertices with adjacency matrix A and let $\boldsymbol{P} = (p_{ij})_{i,j=1,\dots,n}$ be the symmetric $(n \times n)$ matrix of connection probabilities with $p_{ij} \in [0, 1]$. Then, we call G a *heterogeneous Erdös-Rényi graph* $\mathcal{G}(n, \boldsymbol{P})$, if the edges are realizations of stochastically independent Bernoulli random variables. That is, for $1 \leq i < j \leq n$, we have $A_{ij} \sim \text{Bin}(1, p_{ij})$ with $A_{ji} := A_{ij}$, and $A_{ii} := 0$ for all i.

Obviously, this heterogenous model might mimic real-world dynamics better as it gives the possibility for an individual edge probability for each edge. Nevertheless, this results in a very complex model with up to $\frac{n(n+1)}{2}$ edge probability parameters even for a simple graph. Thus, parameter estimation is hardly feasible. To make a heterogenous model better handable, we can consider special cases which put further restrictions on it by grouping similar edge probabilities.

Resulting is a so-called stochastic block model that splits the vertex set into k different blocks $B_1, ..., B_k$ and assigns connection probabilities $p_{s,t}; s, t = 1, ..., k$ with $p_{s,t} \in [0, 1]$ for edges between vertices of blocks B_s and B_t. This special case of a heterogenoues Erdös-Rényi model reduces the complexity to $\frac{k(k+1)}{2}$ edge probability parameters. Intuitively, this allocation makes sense and has great practical potential as vertices of the same block are assumed to have a higher connection probability than to vertices of another block. For the hospital data example, the block structure seems to be naturally given by the status of the nodes (patient, nurse, medical doctors, administrative staff). However, there might be reasons why this allocation is doubtful, e.g. some patients require more care than others resulting in considerably varying edge frequencies for this block to e.g. the block of nurses.

The goodness-of-fit tests proposed in Chapter 4 provide a way of answering such questions and making statistically valid decisions about network modelling. It is not a single test that is proposed, but rather a whole class of tests by deriving limiting distributions for a broad class of network metrics. This is in line with the focus of this thesis to provide flexibility for a wide range of applications as well as making the results of Chapter 2 and Chapter 3 directly usable. Furthermore, a parametric bootstrap scheme for the tests is provided in order to achieve finite sample improvements and to make the tests feasible for highly complex network metrics for which there does no closed form of their moments exist yet.

1.2.3 Related questions in text data

While network data is inherently complex and requires sophisticated analysis methods, this is also the case for other types of complex data. One example is text data. An important task in this area is the detection of change points in order to track topics over time and better understand the development of narratives. The modern problem of detecting fake news also falls within this framework. The challenges are similar to those of monitoring networks, e.g.

1. How can text data be measured, and what are the limitations and possibilities of change detection?

2. How does a change in text data actually look like and can we categorize different types of change?

3. How can the results be interpreted and visualised?

Question 1 is addressed in Chapter 5, where an online changepoint detection scheme for text data is proposed. It is based on a similarity metric and constructed using a rolling window approach for Latent Dirichlet Allocation (LDA). The results are then further processed in Chapter 6, where a rich visualisation approach for this method is developed, together with

a categorisation scheme of potential change types (Question 2), in order to make the results handily interpretable (Question 3).

1.2.4 Structure

Each of the following chapters consists of published and peer-reviewed articles and has been reused with the permission of the copyright holder. Each article is designed to be self-contained and therefore has its own introduction, conclusion and literature review. At the beginning of each chapter, the first page of the original article is reproduced in the format and style of the corresponding journal. This is followed by the entire article in the format of this thesis. Instead of a separate reference list for each chapter, a complete bibliography for the entire dissertation is provided at the end of this thesis. As outlined above, each chapter makes its own contribution to the literature. Nevertheless, the first three chapters contribute to the same field of study and partly build on each other, which is why there is a small overlap of basic concepts between these chapters. The same applies to the last two chapters.

2

Change detection in dynamic networks using network characteristics

Based on: Flossdorf, Jonathan & Carsten Jentsch (2021): "Change detection in dynamic networks using network characteristics", In: *IEEE Transactions on Signal and Information Processing over Networks* 7, pp. 451–464, DOI: 10.1109/TSIPN.2021.3094900.

Abstract

In recent years, the use of dynamic networks became increasingly popular. An important task is to identify differences at particular time points, e.g. for online monitoring, change point detection or testing procedures. Due to the complexity of network data, the statistical analysis is challenging. Therefore, it is usually a main step to characterize the networks by one or few scalar-valued metrics at each time point. As the reduction to such metrics can result in information loss, the understanding of their behaviour in various change scenarios is crucial. However, existing studies commonly use specific data examples and do not give any deeper theoretical insights of the general performances. In this paper, we propose a categorization of different types of changes which can occur in network data. We analyze the suitability and limitations of common network metrics in such situations and give comprehensive explanations of their behaviour. This leads to a well-founded advice of which metrics to use in various application scenarios. Our findings are underlined by an extensive simulation study and some real-world data which involve both time-dependent and independent setups for online monitoring.

Flossdorf, Jonathan & Carsten Jentsch (2021): "Change detection in dynamic networks using network characteristics", In: *IEEE Transactions on Signal and Information Processing over Networks* 7, pp. 451–464, DOI: 10.1109/TSIPN.2021.3094900.

2.1 Introduction

Due to technical progress in recent decades, there is more and more data available which is for example generated by computers, smartphones or sensors. These data provide the basis for further statistical analysis, decisions, and, therefore, for the explanation and understanding of a specific problem. However, not only the amount of data is growing but also its complexity. One might be interested in the properties of observed objects and whether they are related to other objects in a specific manner. This leads to the field of statistical network data. In the most simple form a network consists of a set of nodes (i.e. the observed objects) and a set of edges. An edge is drawn between two objects if they are related in some sense. By transferring this principle to all pairs of objects, a network graph can be created which gives us a good idea of the relationship structure over the whole network.

There are various applications in practice which include e.g. transport [Hulsermann et al., 2004; Wang et al., 2011], social [Huberman et al., 2008; Jackson, 2011] or production [Coe et al., 2008] networks. Typically, a network is in constant change - for example new edges or nodes may enter the network with evolving time. Likewise, existing nodes or edges can be removed at some time points. A single, static network is not able to capture this time component appropriately. Therefore, so-called dynamic networks are used for which the observation window is typically splitted into time stamps at which snapshots of the network are taken. In this context, a popular research topic is the field of change point detection where the main goal is to detect a time stamp at which the network has significantly changed and to analyze what kind of change has happened. A typical approach is to fully observe the process first and to analyze its possible change-points afterwards. This is known as offline change point detection. Online change point detection (also called online monitoring), on the other hand, aims to detect a change as soon as possible after its appearance to rapidly react to the deviation of the in-control state.

Due to the complexity of networks, a direct transfer of traditional change detection approaches [Basseville and Nikiforov, 1993; Montgomery, 2007] is not possible. It is therefore necessary to reduce the time series of networks in terms of complexity. A main approach is to fit a dynamic network model. This includes e.g. GHRG (Generalized Hierarchical Random Graphs), where two versions of the model are examined - one represents the hypothesis of no change and the other a change of parameters at a particular time point [Peel and Clauset, 2015]. A testing procedure is then applied to check which model is more suitable. Another approach [Cheung et al., 2020] assumes that each individual network follows a Stochastic Block Model (SBM), and an objective criterion based on the Minimum Description Length Principle [Rissanen, 1989] is derived. Similarly, the algorithm SCOUT [Hulovatyy and Milenković, 2016] utilizes information criteria derived from the AIC or BIC. As for online change point detection, the model-based approach is typically performed by monitoring the model parameters over time with the use of traditional control charts like CUSUM or EWMA. Different kinds of models

have been applied which include dynamic versions of degree-corrected stochastic blockmodels [Wilson et al., 2019], temporal exponential random graph models [Malinovskaya and Otto, 2021], and Poisson regression models [Farahani et al., 2017]. However, the reduction to a specific family of models is often questionable. If the data does not fit, these approaches are likely to achieve low performances. The complexity of networks and the consideration of the time component additionally affect that most available models are only able to describe quite specific structures. Furthermore, assumptions like a fixed node set have to be fulfilled.

In this paper, we focus on a more flexible approach, where each network is characterized by a network metric. Hence, we obtain a time series of scalar values which can be monitored with traditional techniques of Statistical Process Control (SPC) [Montgomery, 2007; Oakland and Oakland, 2018]. Some approaches take matrix norms of the adjacency matrices into account, e.g. the spectral norm is used in [Chen et al., 2021] to construct an offline change point detection scheme. Furthermore, the Frobenius norm is used to monitor stock markets across the globe [Banerjee and Guhathakurta, 2020]. In [Barnett and Onnela, 2016], a comparison of different matrix norms for change detection in correlation networks is performed. Further typical characteristics are centrality metrics, such as degree or eigenvector centrality, which were e.g applied to monitor military networks [McCulloh and Carley, 2011]. A few metrics were part of a comparison study for community changes in a dynamic version of the SBM [Yu et al., 2022]. However, existing works missed to give a comprehensive mathematical interpretation of the behaviour of the metrics in the applied situations. Moreover, most studies used quite specific data examples or small simulations where only particular network structures and types of changes are addressed. Given that the advantage of the metric-based approach hugely lies in its simplicity and flexibility, it is crucial to understand the suitability of possible metrics in more general situations. Furthermore, the reduction step to a network characteristic is likely to result in some information loss. Consequently, it is even more essential to identify their strengths and weaknesses to choose a suitable one which captures as much relevant information as possible.

Therefore, this paper aims to systematically analyze the performances of common network metrics in various change situations. It is organized as follows: Section 2.2 contains some notation and formalization settings. In Section 2.3, we propose a general categorization of network changes and give real world examples where each situation is likely to occur. In Section 2.4, we interpret the monitoring performance of each considered metric and give explanations for their suitability to detect certain types of changes. These findings are subsequently underlined and further intensified by the use of a simulation study and a real-world example in Sections 2.5 and 2.6, respectively. Throughout the paper, we focus on online monitoring, but the results can be used for offline change point detection and the construction of testing procedures (e.g. two-sample tests) as well. The final Section 2.7 consists of some concluding remarks.

2.2 Setting

Let us denote a single network (also called graph) with G. It consists of a set of nodes V and a set of edges E (also called links). We denote the number of nodes as n and the number of links as m. Furthermore, the network can be represented by an adjacency matrix A of dimension $n \times n$. In Section 2.4, we mainly focus on unweighted and undirected networks (i.e. binary and symmetric adjacency matrices) such that a matrix element $a_{ij} = 1$ represents a relationship between nodes i and j and $a_{ij} = 0$ otherwise. However, most of the results can be transferred to directed and weighted networks as well. Self-loops, i.e. $a_{ii} = 1$, may be included. We denote a dynamic network $\mathcal{D}$ as a sequence of networks of length T such that $\mathcal{D} = \{G^{(t)}, t = 1, \dots, T\}$. The corresponding sequence of adjacency matrices is denoted by $\{A^{(t)}, t = 1, \dots, T\}$.

Although our results can mostly be utilized for offline change point detection and testing procedures as well, we mainly focus on online monitoring if not stated otherwise. Hence, the goal is not only to detect a change as soon as possible but also with a predetermined false alarm rate to achieve reliable results. In this context, the general idea of SPC can be applied [Montgomery, 2007]. To do so, a typical in-control state, in which no meaningful variation of the system is present, needs to be defined (Phase I). The goal at each newly observed time point is then to check whether the network is still in its in-control state or whether some unusual deviation occurred (Phase II). Hence, a sequential statistical test with the following pair of hypothesis is applied [Malinovskaya and Otto, 2021]:

$$H_0^{(t)} = \text{Network at time } t \text{ is in-control} \quad \text{vs.}$$
$$H_1^{(t)} = \text{Network deviates from the in-control state at time } t.$$

Note that this procedure is usually summarized in control charts which offer an understandable visualization of the network behaviour at the examined time points. There exist several different possibilites to construct suitable control charts. Popular examples are Shewhart charts, where the control statistic is directly monitored in a memory-free manner, and moving average charts like EWMA for which past observations are included. Furthermore, different approaches to calculate the control limits can be applied which especially involve distribution assumptions or distribution-free methods (e.g. bootstrap). Sensible analysis of the suitability of different construction principles of control charts for various situations is beyond the scope of this work. To achieve reliable results, we use common approaches [Montgomery, 2007] to construct control charts for network data which fit our purposes in Sections 2.5 and 2.6.

2.3 Categorization of Changes in Dynamic Networks

In traditional SPC approaches, a time series of scalar values is monitored which means that the detection is mostly limited to a shift of the mean or variance of the used metric [Montgomery, 2007]. Furthermore, the interpretation is often not trivial, because this shift usually does not hint to the reasons why some anomalous behaviour occured. In this regard, the use of network data not only leads to a more informative representation of an underlying system but also to a better interpretation of change-points. However, the utilization of more information obviously makes the task of finding appropriate statistical analysis concepts more challenging, e.g. it is not trivial how a change point in dynamic networks looks like, because there is not something obvious like a shift of the mean anymore. In fact, we have to consider all structural elements (e.g. links, nodes) of a network to ensure that we capture the majority of possible changes. Based on this, we propose a general categorization of change-types in dynamic networks. Every case takes one particular structural network element into account and describes possible changes which can be caused by it. For interpretation purposes, we assume that all other structural elements than the one examined stay more or less stable and do not influence the change. However, we address mixed-type changes, where some or all elements can cause changes simultaneously, in a special case. Additionally, within every case, a more specific distinction between global changes, which affect the whole network evenly, and local changes, which only affect parts of the network, is carried out.

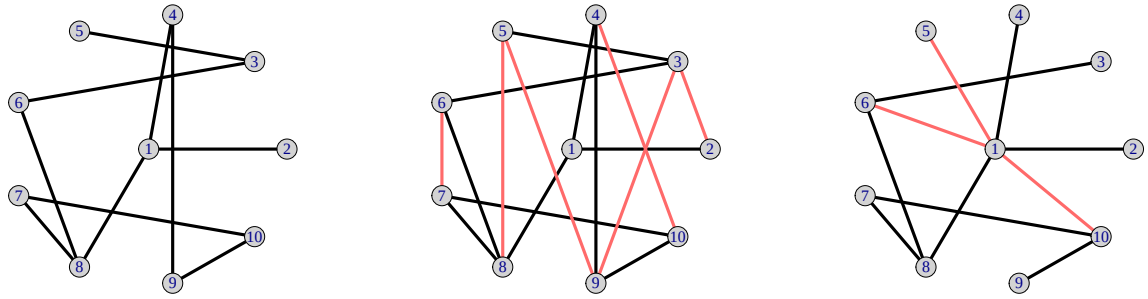
2.3.1 Link Changes

One structural element of networks are links which relate nodes with respect to a particular property.

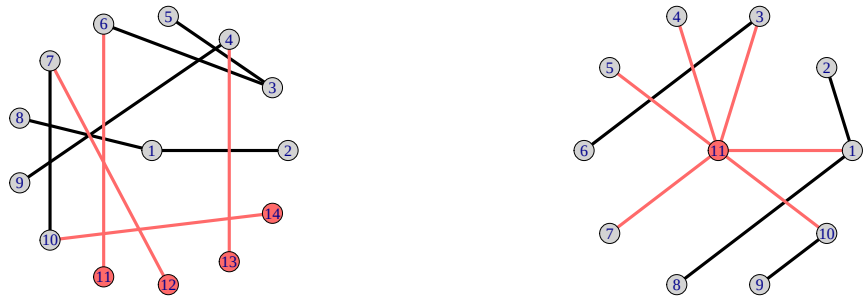
Global Link Changes (GLC)

The overall link amount triggers the change as it is either significantly increased or decreased compared to the in-control situation. The whole network is quite evenly affected such that the general network structure (e.g. hierarchy) stays relatively stable. For illustration purposes, a minimal example is given in Figure 5b, where the link amount increased compared to a representative network of the in-control state of Figure 5a which results in a more dense network. The new links were spread out evenly which means that the structure stays stable.

Examples for these situations are cyber-networks [Morales et al., 2010] where an increased communication between virtual actors implies a possible malware in the system, or social



(a) in-Control **(b)** GLC, A.1) **(c)** LLC, A.2)



(d) GNC, B.1) **(e)** LNC, B.2)

Figure 5: Examples for types of changes in network data. (a) represents the in-control state and (b)-(e) possible changes. Deviations from the example of the in-control state are highlighted in red.

networks with suddenly increased or decreased communication between all actors, e.g. terrorist networks where this may hint to a possible attack in the near future [McCulloh and Carley, 2011].

Local Link Changes (LLC)

These situations are usually characterized by a slighter link amount change, because only a few specific nodes are addressed. This uneven distribution is then likely to produce structural changes, because these nodes gain or lose dominance compared to others in terms of relationships. An example is shown in Figure 5c. While the pure amount does not change as much as for GLCs, the new links were spread out unevenly with a main focus on node ①. This triggers a structural change, because the network moved from a flat hierarchy, in which every object has approximately the same number of relationships, to a more strict hierarchy with node ① as the hub.

Typical application fields are social networks with hierarchy or community changes (e.g. political networks). Another example is the formation of hotspots like in virus networks (regional outbreak of a disease) [Forster et al., 2020; Paull et al., 2012] or tourist networks (the most trendy and popular destinations) [Wu et al., 2018].

2.3.2 Node Changes

Another structural element of networks are nodes which stand representative for the different objects in the system.

Global Node Changes (GNC)

This type of change is triggered by a significant change of the mean amount of nodes which means that either new nodes enter or existing nodes leave the system. In the former case, the new nodes are assumed to be with equal rights to the existing ones such that their entrance in the network does not trigger heavy structural changes. The same applies for nodes leaving the network. A minimal example is illustrated in Figure 5d where the node amount increased by 40%.

This type of change often occurs in friendship networks. If somebody is admitted into a new organization (e.g. sports club) or is moving into another city, he/she will find more friends while still keeping the majority of his/her old ones. Another example is the effect of a new advertising strategy which acquires new customers. Likewise, we can observe it in production networks by adding new machines (removing old, unnecessary machines) or in transport networks [Rodriguez-Nunez and Garcia-Palomares, 2014] by connecting locations of a new region.

Local Node Changes (LNC)

These changes take the properties of the new or removed nodes into account. The pure amount is not important, it is more about a small number of nodes entering or leaving the network which highly influence its structure. A typical example is the entrance of a few central nodes which kind of gather many links (i.e. replacing old ones which are not longer necessary). This behaviour is illustrated in Figure 5e where the entrance of node ⑪ pushes the network towards a more centralized layout.

Examples of such changes are the invention of more efficient machines in production networks or the creation of new leading positions in profession networks. Furthermore, new central locations may be introduced in a supply chain [Wen et al., 2013].

2.3.3 Extra Information Changes

A basic statistical network captures nodes and links. Hence, the changes described above are the most common ones. However, a huge advantage of network data is the possibility to extend it with all kind of extra information. Thus, we can describe the nodes more explicit by stating properties (e.g. gender in social networks) or individual ID numbers with which they can be distinguished from each other. Additionally, we can specify information about links more precisely, e.g. by stating their intensity (weighted graphs). Change-points can be triggered by a significant change of extra information as well which is again possible for the whole network or for particular parts. Note that such changes are quite specific, since available information varies from one application field to another. Hence, we dispense with a general categorization in this case.

2.3.4 Mixed-type changes

In the cases discussed before, we addressed changes which were triggered by just one type of structural element while the other ones were assumed to stay stable. This distinction is necessary for interpretation purposes in the remainder of this paper. In practice, however, it is also possible that the described cases occur simultaneously or even influence each other, e.g. it is likely that the entrance of new nodes triggers an increased link amount. It is then possible to detect a change point by taking both types of changes into account.

Example: Community Changes

The above categorization is quite general and covers the majority of changes which can occur in network data. This also involves more specific setups such as community changes. Major amount changes of some communities in terms of communication or members can be viewed as GLCs or GNCs, respectively. Community changes from a more structural perspective without rigorous amount changes can be assigned to multiple local changes like LLCs or LNCs. In some rare scenarios there might even be a removal of several nodes which is followed by the addition of other nodes with a similar link pattern as the old ones. In fact, this is a change in the membership of nodes and can therefore be classified as a change in extra information (2.3.3). Further community change scenarios and their linkage to the proposed categorization

are given in Section 2.5.

The idea of this categorization is now used to analyze the impact of various situations on the behaviour of the metrics from a mathematical point of view. In the following section, link and node changes will be addressed - both with the local and global setup. Afterwards, we underline our statements with a simulation study where mixed-type and community changes are addressed as well. The paper aims to gain fundamental insights of the behaviour of the metrics in general change situations. Based on this, most results can be transferred to more specific structures. Due to the large variety of changes, which are triggered by extra-information (2.3.3), and as their nature highly depends on the application field, we will not take these specially into account. However, we point out that such changes can usually not be detected by the considered metrics, since they do not take any kind of extra information into account. In this case, the usage of model-based approaches [Farahani et al., 2017; Malinovskaya and Otto, 2021; Wilson et al., 2019] or similarity measures [Koutra et al., 2016] are more appropriate. However, these approaches are less flexible as they require strict assumptions on the network (e.g. fixed node set).

2.4 Monitoring Ability of Network Metrics

We now move on to the interpretation of common network metrics in the described change situations. For interpretation purposes, we assume that the dynamic network is undirected and unweighted. However, large parts of the results can be transferred to directed and weighted networks. An overview of the considered metrics is given in Table 2.

2.4.1 Matrix-based Metrics

Firstly, we examine the performance of metrics which are directly calculated on the underlying adjacency matrices. This involves matrix norms like the Frobenius norm or spectral norm. The former is applied in correlation and financial networks [Banerjee and Guhathakurta, 2020; Barnett and Onnela, 2016], and the latter is e.g. used for testing procedures based on which a change detection scheme is created [Chen et al., 2021]. Moreover, eigenvalues can be used for network comparisons [Ghoshdastidar and Luxburg, 2018]. In this regard, we show that the largest eigenvalue corresponds to the spectral norm for undirected graphs.

Table 2: Considered types of metrics in this work.

Type	Metric	Abbreviation
Matrix-based	Frobenius norm	F
	spectral norm	S
	largest eigenvalue	LEi
Centralities	average closeness	C_m
	closeness deviation	C_d
	average degree	D_m
	degree deviation	D_d
	average betweenness	B_m
	betweenness deviation	B_D
	average eigenvector	E_m
	eigenvector deviation	E_d
Other	average path length	AvP
	transitivity	Tr

Frobenius Norm

The Frobenius norm is a popular metric to summarize a matrix to a scalar value and is often used to compare different matrices in terms of their similarity. For a given matrix $A \in \mathbb{R}^{n\times m}$, we define its Frobenius norm by

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2}.$$

In the case of binary adjacency matrices (i.e. unweighted networks), it applies $a_{ij} \in \{0, 1\}$. Consequently, it holds $a_{ij}^2 = a_{ij}$ and the metric is simply calculated by the root of the number of 1's in A. Note that this is the root of twice the number of links in an undirected network.

Therefore, it only makes sense to use it, if the change is caused by differences in the amount of links which especially involve GLC situations. In many practical relevant LLC cases, the metric is likely to have a reasonable performance too. Nevertheless, the amount changes are noticeably slighter and the more decisive structural changes are not captured which means that the metric is not as effective. There are even LLC scenarios which the Frobenius norm is not able to detect at all, e.g. if multiple local changes occur which rule each other out in terms of link amount.

If the link amount stays more or less stable and is not the reason for a change, then the metric is inappropriate for monitoring purposes. Changes which are purely caused by nodes, such as GNC or LNC setups, are ignored.

Spectral Norm / Largest Eigenvalue

Another possibility to compare matrices, is given by the calculation of the spectral norm which is defined by

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 .$$

It can be simplified as follows: The spectral norm corresponds to the largest singular value of A which is the root of the largest eigenvalue of A^TA. In our assumed case of an undirected network, the adjacency matrix A is symmetric. This leads to $A = A^T$ and $A^TA = AA$. Hence, the spectral norm corresponds to the root of the largest eigenvalue of $AA = A^2$.

Let us now denote q as the largest eigenvalue and x as the corresponding eigenvector of A which both follow the definition $Ax = qx$. Then, there is a connection to A^2 via

$$A^2x = A(Ax) = A(qx) = q(Ax) = q(qx) = q^2x.$$

Consequently, q^2 is the largest eigenvalue of A^2 and $\sqrt{q^2} = q$ is the largest singular value of A. Putting all together, the spectral norm corresponds to the largest eigenvalue of the underlying symmetric adjacency matrix A.

In this context, it applies $\bar{d} \leq q \leq d_{max}$ [Brouwer and Haemers, 2011], where $\bar{d}$ and d_{max} are the average and maximum degree, respectively. Hence, the spectral norm depends on the degree structure of the graph and our simulations in Section 2.5 show that it tends to detect a change, if the upper and lower bounds are affected and move into the same direction. If the bounds are not affected, the metric does not change much either. A problem can occur, if the interval is stretching in both directions such that the upper bound increases and the lower bound decreases. In these situations it is likely that the metric is also not much affected by the change.

Therefore, we will examine the behaviour of the bounds $\bar{d}$ and d_{max} to make a statement about the performance of the spectral norm. In this regard, note that $\bar{d} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}$ only captures amount changes. Obviously, it addresses links with the summation term and nodes with the prefactor. On the other hand, d_{max} can be affected by structural changes. Especially layout changes to a more centralized network with a dominant node, would lead to a higher value.

Consequently, global changes, which especially address link or node amount, are likely to be detected, as the bounds move into the same direction. An increased link amount would for example lead to increased values of both d_{max} and $\bar{d}$, whereas an increased node amount would result in decreased values, because the network becomes more sparse.

The performance for local changes is more dependent on the particular situation. It is possible to detect those changes in which the structural change has the same effect on d_{max} as the

Table 3: Behaviour of the spectral norm and their bounds in the example situations of Figure 5.

Change-Type	$\bar{d}$	d_{max}	S
GLC	$\uparrow$	$\nearrow$	$\uparrow$
LLC	$\nearrow$	$\uparrow$	$\nearrow$
GNC	$\downarrow$	$\searrow$	$\downarrow$
LNC	$\searrow$	$\nearrow$	$\rightarrow$

$\uparrow, \downarrow$: large positive/negative impact
$\nearrow, \searrow$:slight positive/negative impact
$\rightarrow$: no relevant impact

amount change has on $\bar{d}$. The LLC case illustrated in Figure 5c provides such an example, because the slightly increased link amount leads to a higher value of $\bar{d}$ and the more centralized structure to an increased d_{max}. Nevertheless, we can observe contrary behaviour as well. In the LNC example of Figure 5e, the increased node amount results in a lower value of $\bar{d}$, whereas d_{max} takes a higher value due to the more centralized structure. As a result, both bounds pull into opposite directions which means that the effect on the spectral norm tends to get lesser. See Table 3 for a summary of the behaviour of the bounds for these example situations.

2.4.2 Centrality Metrics

The goal of network centrality metrics is to assign a score to each node which gives information about its individual influence within the network [Kolaczyk and Csárdi, 2014; Newman, 2018]. By a meaningful summary of these scores, we obtain a global centrality score for the whole network. One possibility is to calculate the node-wise average. If $c(v)$ denotes the centrality score of node v, we can express this formally by

$$c_{\text{avg}} = \frac{1}{n} \sum_{v \in V} c(v).$$

Furthermore, we can emphasize the score of the most central node by calculating the deviation of all scores to it

$$c_{\text{dev}} = \sum_{v \in V} (\max_{u \in V}(c(u)) - c(v)).$$

This version results in lower values if the network is homogenous (i.e. the scores of the nodes do not differ much), and higher values if the network is more unbalanced.

Note that various criteria can be used to measure the influence of a node. We will interpret the two presented centrality versions for those criteria which are mostly used for monitoring [McCulloh and Carley, 2011; Yu et al., 2022].

Closeness Centrality

A popular centrality criterion is the closeness which is formally given by

$$c(v) = \frac{1}{\frac{1}{n-1} \sum_{u \in V} \mathrm{d}(v,u)},$$

where $\mathrm{d}(v,u)$ denotes the shortest distance between nodes v and u. A node achieves higher scores the better it is connected to all other nodes because its inverted average of shortest distances is considered.

Average Closeness

Consequently, the average gets higher if the paths between the nodes become shorter. Hence, the metric perfoms reasonable in terms of global changes which have an impact on the density of the network. An overall increase/decrease of links trivially causes shorter/longer paths throughout the whole network and therefore higher/lower scores of the average closeness. Adding new nodes to the network would result in more sparse networks and longer paths. Analogously, the removal of nodes tends to result in denser networks and shorter paths.

Regarding local changes, the metric behaves differently. Here, the amount changes are not as drastic which is why they have a slighter impact. Much more, the changed structure has to be considered. A rather heterogenous network with a few dominant nodes tends to generate shorter paths than more homogenous networks, because most nodes are indirectly related through the dominant one(s). This especially applies for rather sparse networks, whereas the effect is weakened when the network is highly dense and direct connections are likely. Consequently, it is possible to detect structural changes, but the impact on the metric is often quite low. Bear in mind that slight amount changes of nodes/links can rule this impact out, if they have an opposite effect on the metric (e.g. the adding of a few dominant nodes).

Closeness Deviation

Regarding the deviation metric, we obtain different performance results. It takes lower values if the nodes are relatively equi-central and higher values if there are a few dominant nodes in terms of closeness. Consequently, the metric has problems when it comes to global changes which affect the whole network more or less evenly. Pure amount changes tend to have the same impact on the centrality score of each node. In the GLC situation of Figure 5b for example, each score increases more or less evenly due to the more dense network, while it's the other way round for the GNC example in Figure 5d. Hence, the deviation is barely affected and the metric is usually not able to detect these type of changes.

Obviously, the metric performs better when it comes to local changes. For example, a centralized layout with the formation of some hub nodes results in a better connection and shorter paths for these nodes which means that the maximum score increases noticeably. Bear in mind that the scores of the other nodes tend to improve as well due to the more central layout (as described above). However, this effect is slighter than on the maximum score and, thus, the impact on the deviation is usually noticeable.

A general issue, when dealing with deviation metrics, is that we often observe high variances in the in-control state, because its calculation is quite dependent on a single value (i.e. the maximum score) which can easily be influenced by small in-control changes or noise. Consequently, the control limits often have to be set wider which has a negative impact on the effectiveness.

Degree Centrality

The degree centrality score of a node is simply given by its degree in the network, i.e. the amount of relationships which the node shares with all other nodes.

Average Degree

We already encountered the average degree $\bar{d}$ as the lower bound of the spectral norm. In GLC situations, it is generally a bit more effective than other centrality criteria, because the changed link amount is directly captured. The effect of LLCs, on the other hand, is noticeably slighter. Structural changes cannot be detected, because the smoothing with the average only considers the pure link amount and not the fact in which manner the links are spread over the network. Hence, similarly to the behaviour of the Frobenius norm, these changes can only be detected if the overall link amount changes noticeably as well. Note that multiple local changes, which cancel each other out in this regard, are ignored by the metric.

Contrary to the Frobenius norm, the metric is able to detect some node changes which are considered by dividing the sum of degrees with the number of nodes n. However, this is limited on GNCs for which the link amount is assumed to stay stable. Mixed-type changes, on the other hand, can cause problems when both numerator (sum of degrees) and denominator (number of nodes) change in the same direction. The interpretation of LNCs is analogous to the one of LLCs.

Degree Deviation

Whenever there is a volatile change of the maximum degree compared to the other degrees, e.g. in hierarchy changes, the monitoring performance is reasonable. Consequently, the metric

is in general able to detect most changes of the LLC and LNC type. Analogously to the closeness deviation however, this metric is not flexible, as global changes are mostly ignored and robustness issues may occur.

Betweenness Centrality

Another popular centrality criterion is based on betweenness which determines the fraction of shortest paths between all pairs of nodes which pass through the node in question. Subsequently, the sum of these fractions results in a centrality metric for each node. We can formally represent this by

$$c(v) = \sum_{v \neq u \neq w} \frac{\sigma_{uw}(v)}{\sigma_{uw}},$$

where σ_{uw} is the total amount of shortest paths between nodes u and w. Furthermore, $\sigma_{uw}(v)$ denotes the number of these paths which pass through the node in question v. Consequently, a node achieves higher scores the more it is involved in shortest paths between all pairs of other nodes.

Average Betweenness

Whenever the network becomes more sparse, the shortest paths between nodes get longer. Hence, there are usually more nodes to be passed through which results in overall higher centrality scores (due to the increased $\sigma_{uw}(v)$) and, therefore, in a higher value of the average betweenness. Vice versa, the metric decreases if the network gets more dense, because shortest paths get shorter and the nodes are more directly connected with less other nodes to be passed through. Consequently, the metric is usually able to handle global changes which affects the overall density of the network. This is the case for both GNCs and GLCs.

The interpretation of local changes is more difficult. An advantage of the betweenness is that structural changes do have an impact on the average which especially was not the case for the degree centrality. To make this point more clear, let us consider the example of the movement from a flat network hierarchy to a more centralized layout. Most of the paths are then connected through hub nodes which result in a volatile increasement of their centrality scores, while there is also likely to be a slight increasement for other nodes which profit of the proximity to the hub nodes. Only a few outliers, which are not directly connected with them, are losing in terms of their centrality score. Overall, the centralization has a slight but likely identifiable effect on the metric, if the change is heavy enough. Note that this effect depends on the density of the network, as such layout changes are more effective for more sparse networks. In the LLC and LNC scenarios, there is also a small change in the amount of

links or nodes possible. If this has the same effect as the structural change, then the metric is likely to perform reasonable. Problems can obviously occur if the structural effect on the metric is contrary to the one of the amount change.

Betweenness Deviation

This metric behaves quite similarly to its counterparts of the other criteria. If there are global changes, which concern the whole network equally, then the centrality scores of each node change quite evenly. Hence, the deviation to the maximum score stays more or less stable and the metric is in most cases not able to detect the change.

For the local setup, it is different. If we pick up the example of the hierarchy movement from above again, then we already stated that the scores of the dominant nodes increase drastically, whereas the scores of the other nodes change more slightly. Thus, the impact on the calculated deviation is noticeable. If we consider the opposite case, in which we move from a more strict hierarchy to a more flat hierarchy, the effect is vice versa. The pure addition or removal of nodes or links does not affect this behaviour much as explained above. Nevertheless, the metric suffers under similar robustness issues as its counterparts.

Eigenvector Centrality

Eigenvector centrality can be interpreted as a natural extension of the degree centrality. It is measured relatively to the amount of relations a node has to other nodes, and relationships to high-scoring nodes are more valuable than these to low-scoring ones. Therefore, not only the number of connections is taken into account but also the influences of the nodes to which the connections exist. By denoting the score of node i with x_i, we can define the eigenvector centrality formally by

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} a_{ij}x_j,$$

where λ is a constant. If we define the vector of scores as $x = (x_1, x_2, \ldots)$, it is possible to express the equation in matrix form

$$\lambda x = Ax.$$

Hence, λ is an eigenvalue of A and x is the corresponding eigenvector which includes the centrality scores. For better interpretation, the scores should all have positive signs. In this context, the Perron-Frobenius theorem guarantees that the eigenvector of the largest eigenvalue does exactly meet this condition [Brouwer and Haemers, 2011].

Putting all together, the eigenvector centrality corresponds to the largest eigenvector of the adjacency matrix. The eigenvector values are scaled to be able to compare the different scores properly with each other. In the usual network context, the scaling is done using the maximum score to transform the values to the $[0, 1]$ interval. Alternatively, we can just use the internal scaling of the eigenvector, as it has unit length in the euclidean norm. Both approaches could be used for monitoring purposes and our simulations show that they behave quite similarly to each other. We use the former, because it is more common in the field of network analysis.

In this context, at least one node achieves the maximum centrality of 1. Therefore, the eigenvector deviation simplifies to:

$$c_{\text{dev}} = \sum_{v \in V} (\max_u(c(u)) - c(v)) = \sum_{v \in V} (1 - c(v)).$$

Hence, it is obvious that c_{dev} is an affine linear transformation of c_{avg}, i.e. both metrics achieve the same monitoring performances. We will therefore only focus on the average version in the following.

Average Eigenvector/Eigenvector Deviation

If there are many nodes, which are as influential or nearly as influential as the node with the largest eigenvector score, then their centrality scores take high values as well. For example, a network layout like a ring will produce scores of 1 for each node, because every node has the same impact. Layouts like a star, on the other hand, will result in low centrality scores for all nodes except for the central one, because their influence is very low compared to its. Hence, larger scores are usually created by more flat hierarchies and lower scores by more central layouts which are mainly focused on a few nodes. Consequently, the average eigenvector is kind of predestined for detecting structural changes like LLCs and LNCs. This is because the sum of the individual scores is directly used which means that the metric takes higher values the more equal the influence of all nodes is.

Global changes, on the other hand, do not have a direct impact on the average eigenvector, because they are assumed to not change the structure of a network. Nevertheless, we can observe an indirect effect of GLCs on the metric, because the probability of creating a flat network hierarchy increases with the network's density. Note that for this principle to apply, a random distribution of the links has to be assumed (like in GLCs). For example, consider a network with ten nodes and ten links. There are only a few possibilities to create a rather flat hierarchy and only one to create complete equality. On the other hand, the great majority of possibilities would lead to very heterogenous scenarios. If we observe 36 links instead, then the probability to create a flat hierarchy increases drastically, because there are far less possibilities for a rather central layout. Hence, GLCs can be detected but only if they are heavy enough.

For GNC scenarios, adding new nodes (and therefore new centrality scores) to the network, while keeping the general structure, leads to higher values of the sum of scores. However, this effect is ruled out, because we consider the average as we multiply the sum with $\frac{1}{n}$. Therefore, we can only observe a similar impact as before. The removal of nodes results in a more dense network in which the sum of centrality scores takes higher values on average. For the addition of nodes, the behaviour is vice versa.

Consequently, the average eigenvector is a quite flexible metric, as it is theoretically affected by all types of considered changes. However, the impact of global changes is rather low, i.e. the metric is not as effective in these situations.

2.4.3 Other Network Metrics

Beside centrality and matrix-based metrics, it is of course possible to use any other meaningful network characteristic. We will examine two of them - the average path length and the transitivity - which both play an important role in so-called Watts-Strogatz models [Watts and Strogatz, 1998] and were already mentioned for change detection [Yu et al., 2022].

Average Path Length

The average path length is calculated by

$$l_{\text{avg}} = \frac{1}{n(n-1)} \sum_{u \neq v} d(u, v).$$

Thus, the mean length of the shortest paths between all pairs of nodes is determined. The concept is quite similar to the average closeness. However, it is a more global metric of the shortest paths in the whole network, whereas the average closeness calculates the node-wise mean of all shortest paths regarding a specific node once at a time.

Despite of these slight differences, the interpretation is very similar. Obviously, GLCs have a direct impact on the metric. An increased amount of links results in a better connected network (i.e. shorter paths), for a decreased amount this is vice versa. Regarding GNCs, more nodes make longer paths, because the existing links need to be shared between an increased amount of nodes, whereas less nodes create shorter paths.

The behaviour in local change scenarios is again similar to the average closeness. For a more central layout, the average path length tends to decrease slightly due to the likeliness that new shortest paths are created which pass through the hub nodes. An additional small link or node amount change towards a more dense network would in this case be supportive. Generally speaking, however, the effectiveness of the metric in such situations is relatively low and

dependent on other data conditions (e.g. density, intensity of the change).

Transitivity

The transitivity is a popular metric to measure the degree of clustering in a network. It is a fundamental part of the concept of the Small-World phenomena which plays an important role in many network topics (e.g. Watts-Strogatz model). It is defined by

$$T_{Cl} = \frac{3 \cdot \text{number of triangles}}{\text{number of triplets}}.$$

A triplet corresponds to three nodes which are connected by either two or three links. A triangle is a special case of a triplet in which all three links are present. The transitivity calculates the fraction of triangular structures to all existing triplets. Therefore, a higher value speaks for a more clustered graph. The correction factor in the numerator is because a triangle includes three triplets, one centered on each of the involved nodes.

Obviously, the metric is suitable for layout changes which result in a different clustering (e.g. formation of subgroups). However, for more general changes an interpretation is more difficult, because the metric is often not directly influenced by the change. How the metric behaves is, therefore, dependent on the in-control state of the network. It is worth noting that the addition of a single triangle has a large effect on the metric, because it corresponds to three triplets. Regarding global changes, the movement from a sparse network to a more connected, but still relatively sparse network, is likely to produce a decreased transitivity, because the formation of connected subgraphs, like triangles, is very unlikely in this case. If there is a change from a quite connected network to an even more connected network, then the probability of the formation of new triangles tends to be higher which would result in increased values of the transitivity.

LLCs and LNCs only have an indirect impact as well. The movement to a more central layout generally means that triangles are likely to occur if two or three hub nodes are involved. For all other node combinations, it is less likely that a triangle is formed due to the low probability of ties between non-central nodes. Nevertheless, this is of course dependent on the particular situation.

2.5 Simulation Study

We now underline our interpretations with a simulation study. As typical in many applications of SPC and network monitoring, we generate the networks independently over time and therefore treat the emerging time series as an independent process. Note that this assumption

Table 4: Data generation processes using Erdös-Renyi (ER) graphs.

Type	Data Generation
GLC	In-Control: ER graph with probability p and n nodes Out of Control: ER graph with probability $\tilde{p}$ and n nodes
LLC	In-Control: ER graph with probability p and n nodes Out of Control: heterogenous ER graph with a changed probability $\tilde{p}$ for k central nodes
GNC	In-Control: ER graph with at each time point a randomly chosen node size in $[n - n \cdot d, n + n \cdot d]$ and a randomly chosen link amount in $[m - m \cdot d, m + m \cdot d]$. Out of Control: ER graph with at each time point a randomly chosen node size in $[\tilde{n} - \tilde{n} \cdot d, \tilde{n} + \tilde{n} \cdot d]$ and a randomly chosen link amount in $[m - m \cdot d, m + m \cdot d]$.
LNC	In-Control: ER graph with probability p and at each time point a random node size $n_{\text{flex}} \in [n - n \cdot d, n + n \cdot d]$. Out of Control: heterogenous ER graph with at each time point a random node size $\tilde{n}_{\text{flex}} \in [\tilde{n} - \tilde{n} \cdot d, \tilde{n} + \tilde{n} \cdot d]$ and changed probability $\tilde{p}$ for k central nodes and a probability of $\frac{pn_{\text{flex}} - \tilde{p}k}{\tilde{n}_{\text{flex}} - k}$ for all other nodes.

is rather critical for dynamic networks in practice. It is e.g. likely that the presence of a link influences its presence at future time points. To address both cases, we will therefore analyze real-world data in Section 2.6 with a time-dependent setup.

2.5.1 Setup

Each of the change-types GLC, LLC, GNC and LNC are represented by three different scenarios with various intensities (small, moderate and heavy). See Table 4 for our used data generating processes, and Table 5 for the associated parameter setup. To achieve reliable results, each scenario is replicated 1000 times. For local changes, we particularly consider structural changes influenced by a slightly increased node or link amount which moves the network from a flat to a more centralized layout. A heavier change intensity is then understood to take place, if the hierarchy becomes more strict such that there are fewer but even more dominant hubs. Other local change cases, for which the metrics perform differently, were emphasized in their interpretations in Section 2.4.

For monitoring, EWMA control charts [Roberts, 1959] are used for which the control statistic is defined as $z_t = \lambda x_t + (1 - \lambda) z_{t-1}$. As we analyze every metric separately, we use univariate charts and, hence, x_t is the observation of the applied metric at time t. We try different values of the smoothing parameter $\lambda \in \{0.1, 0.2, \ldots, 0.5\}$ and report the ones with the best result in each case. The control limits are derived under the assumption of a normal distribution of the control statistic and are set such that an in-control average run length (ARL_0) of 200

Table 5: Parameter setup for each situation according to Table 4 with $n = 100$.

Type	Intensity	Parameter Setup
GLC	small	$p = 0.3$, $\tilde{p} = 0.31$
	moderate	$p = 0.7$, $\tilde{p} = 0.65$
	heavy	$p = 0.4$, $\tilde{p} = 0.5$
LLC	small	$p = 0.2$, $\tilde{p} = 0.3$, $k = 5$
	moderate	$p = 0.4$, $\tilde{p} = 0.55$, $k = 3$
	heavy	$p = 0.3$, $\tilde{p} = 0.6$, $k = 1$
GNC	small	$d = 0.05$, $m = 1500$, $\tilde{n} = 115$
	moderate	$d = 0.05$, $m = 1500$, $\tilde{n} = 130$
	heavy	$d = 0.05$, $m = 1500$, $\tilde{n} = 150$
LNC	small	$d = 0.03$, $p = 0.2$, $\tilde{n} = 105$, $\tilde{p} = 0.3$, $k = 5$
	moderate	$d = 0.03$, $p = 0.7$, $\tilde{n} = 102$, $\tilde{p} = 0.85$, $k = 2$
	heavy	$d = 0.03$, $p = 0.3$, $\tilde{n} = 101$, $\tilde{p} = 0.7$, $k = 1$

time points is maintained. As the EWMA chart is robust to non-normality, its application is suitable for our purposes [Montgomery, 2007]. An alarm is triggered if a value beyond the control limits is observed.

As typical in SPC, the process (i.e. sequence of networks) is subdivided into Phase I and Phase II. In the former, the control chart is calibrated by estimating mean and standard deviation of the underlying metric, whereas in the latter, the actual monitoring is examined. We use the first 1000 networks for Phase I and Phase II starts at $t = 1001$. The change occurs at $t = 1031$. In practice, it might be sometimes unlikely to observe such a long sequence of in-control networks. Hence, we repeated the simulation study for shorter Phase I lengths (not reported), but observed no major differences leading to qualitatively similar results. The performance of each metric is measured by calculating the postchange average run length which is typically denoted as ARL_1. It measures the delay to detection, i.e. the number of time points an alarm is triggered after the actual change has occurred.

2.5.2 Results

It can be seen that the results displayed in Table 6 are in accordance with our interpretations. The usage of the Frobenius norm is critical, as it is only able to detect link amount changes. But even in these situations, the metric is not able to outperform other suitable counterparts. The spectral norm is quite flexible, as it can perform well in both global and local change scenarios. However, under particular circumstances, which were discussed above, it does not work reasonable, like in the LNC examples here. Regarding the centrality metrics, the results underline that a meaningful choice of the used version (average or deviation) is more essential than the choice of a particular criterion. The average version works well for global changes, whereas the deviation is often the better choice for local changes, especially if we observe fewer

Table 6: ARL_1 of the metrics. Best ARL_1 for each scenario is printed in bold.

Type	Intensity	F	S	C_m	C_d	D_m	D_d	B_m	B_d	E_m	Tr	AvP
GLC	small	4.40	**4.25**	4.32	79.91	4.34	96.28	4.33	134.21	131.29	4.98	4.33
	moderate	**1.00**	**1.00**	**1.00**	116.30	**1.00**	34.87	**1.00**	3.26	9.02	1.04	1.03
	heavy	**1.00**	**1.00**	**1.00**	9.66	**1.00**	109.64	**1.00**	3.40	5.65	**1.00**	**1.00**
LLC	small	3.01	**2.49**	3.38	2.71	2.98	2.64	3.57	3.16	3.93	2.96	3.57
	moderate	5.18	4.11	4.64	**1.75**	5.18	1.86	5.18	2.02	2.10	3.58	5.18
	heavy	9.89	5.59	8.09	**1.03**	9.85	**1.03**	9.86	**1.03**	1.04	3.96	9.86
GNC	small	211.21	1.84	1.57	9.20	1.81	17.50	**1.23**	11.99	8.84	1.55	1.52
	moderate	208.17	1.02	**1.00**	5.21	1.03	5.45	**1.00**	6.35	3.01	1.02	**1.00**
	heavy	198.13	**1.00**	**1.00**	11.18	**1.00**	2.84	**1.00**	3.56	1.73	**1.00**	**1.00**
LNC	small	167.96	13.22	1.52	1.36	4.86	1.46	1.30	**1.29**	1.49	3.87	1.50
	moderate	222.60	10.28	1.26	1.30	9.02	1.24	1.39	**1.11**	1.20	1.26	1.23
	heavy	207.50	94.37	14.10	**1.00**	58.22	**1.00**	11.31	**1.00**	**1.00**	15.98	9.63

hub nodes. There are only little differences regarding the different criteria. The degree is a bit more effective for GLCs and GNCs, as it directly takes global changes into account, whereas closeness and betweenness are more flexible, especially for the average version. The average eigenvector is in general affected by all considered changes. For global changes, however, the influence is much slighter such that small intensities cannot be detected. The transitivity is not directly affected by the examined changes and therefore performs quite differently which is also dependent on the general density of the graph as explained above. Consequently, it is the metric which is characterized by the highest variance within the cases. Additionally, the results underline the similarity of average path length and average closeness, as they achieve similar performances.

2.5.3 Special Cases: Community/Mixed-Type Changes

We now illustrate how more specific changes affect the behaviour of the metrics. Therefore, we especially focus on the rather popular subject of community change detection. Hence, we assume the network to follow a block structure which itself changes after some period of time. We therefore examine four different changes with the use of stochastic blokmodels (SBM). Parameter setups of the in-control state and the changes are summarized in Table 7. In this context, K represents the number of communities, p_i the intra-group link probability of group $i \in \{1, ..., K\}$, p_{ij} the inter-group link probability between groups i and j, and n_i the number of nodes in group i. The corresponding results are presented in Table 8. The overall monitoring procedure is the same as before.

The first scenario considers changes with an increased popularity of the network which results in an increased communication within and between all groups and the entry of new members.

Table 7: SBM Paramter Setups for the Community Changes

No.	In-Control	Changes
1	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.8, p_{ij} = 0.1,$ $n_1 = n_2 = 33, n_3 = 34$	$p_1 = 0.9, p_2 = 0.7,$ $p_3 = 0.9, p_{ij} = 0.2,$ $n_1 = n_2 = n_3 = 40$
2	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.8, p_{ij} = 0.1,$ $n_1 = n_2 = 33, n_3 = 34$	$p_1 = 0.4, p_2 = 0.9$
3	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.3, p_{ij} = 0.1,$ $n_1 = 30, n_2 = 20,$ $n_3 = 50$	$K = 4, p_4 = 0.3,$ $n_3 = 30, n_4 = 20$
4	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.3, p_{ij} = 0.1,$ $n_1 = 30, n_2 = 20,$ $n_3 = 50$	$K = 4, p_3 = p_4 = 0.49$ $n_3 = 30, n_4 = 20$

Table 8: ARL_1 for the Community Changes. Best ARL_1 for each scenario is printed in bold.

No.	F	S	C_m	C_d	D_m	D_d	B_m	B_d	E_m	Tr	AvP
1	**1.00**	**1.00**	**1.00**	135.22	**1.00**	39.85	**1.00**	3.25	1.79	1.02	**1.00**
2	203.18	19.81	87.59	146.15	206.49	128.24	198.59	30.02	2.72	**1.13**	193.61
3	1.27	4.51	1.31	3.09	1.29	5.49	**1.21**	5.40	2.53	16.11	**1.21**
4	195.19	135.08	23.54	67.99	197.34	133.65	21.40	73.07	122.51	**4.25**	21.43

This case can be easily linked to the proposed categorization in Section 2.3 as a mixed-type change of GLC and GNC. Although the increased node amount would solely result in a sparser network, the increased link amount is clearly heavier such that we observe more dense networks after the change. Hence, the performances of the metrics are similar to the GLC case with moderate intensity of Table 6.

The second scenario addresses changes where the importance of two groups is shifted such that one group communicates less while another group communicates more. To underline the limitations of the metrics, we balanced the shift to avoid link amount changes. Hence, the Frobenius norm and average centralities do not perform well in this particular example. However, the changes can be viewed as two LLCs which rule each other out in terms of link amount. Additionally, the structural change is quite small as only light probability changes in two groups are examined. As a result, even some deviation centralities are often not able to detect the change. This is not the case for eigenvector centrality which is more sensible in such circumstances than other criteria.

The third scenario captures changes in the amount of communities. In the simulated example, we observe a split of one group into two different groups such that the number of communities increases by one. As we expect the same link probability in both new groups as it was in the

old one, the overall number of links decreases. This is because of the decreased link probability between those nodes which were in the same group before the change and in different ones afterwards. Hence, we can categorize it as a GLC which is also underlined by the performances of the metrics.

The last scenario is similar to the one before. However, to again stress the limitations of the metrics, the probabilities of the new groups are set such that no overall link amount change is visible. Note that this is quite unlikely in practice. Hence, there is only a very light LLC which can not reliably be detected by most of the metrics. Only the transitivity achieves satisfactory results in these circumstances as it directly takes the changed cluster structure into account. This is because the probability of three nodes building a triangle is more likely, since the two new groups are locally more dense than the original group of the in control state (less nodes, higher link probability).

2.6 Empirical Data Example

We now move on to the application of the methods in a real-world data example. This especially serves to further illustrate the behaviour and suitability of the metrics for mixed-type changes and for dependent data. We use the daily flight data between US Airports published by the US Bureau of Transportation Statistics. Here, the airports serve as nodes, and an undirected link is drawn as long as there was at least one flight between two airports. Networks are created daily from January 1st, 2018 to October 31, 2020. Cancelled flights are not considered. Bear in mind that the node set is not fixed, as airports can be closed, newly opened or seasonally inactive.

Due to the global pandemic of COVID-19 and the accompanying restrictions of air traffic, we expect a control chart to identify anomalous behaviour in spring 2020. Figure 6 clarifies that we can observe a changed link as well as a changed node amount at this time.

However, the network structure changed as well. To reduce the number of infections, many direct routes between airports were cancelled which especially affected minor airports. Some of them were even completely shutdowned like the Westchester County Airport in New York. To ensure mobility, the routes of the most busiest airports were tried to be maintained as much as possible. This means that the network is pushed towards a more centralized layout, where the hubs gain even more dominance. This behaviour is illustrated in Figure 7, where it can clearly be seen that the central airports mostly conserved their variety of connections, while the network becomes way more sparse elsewhere.

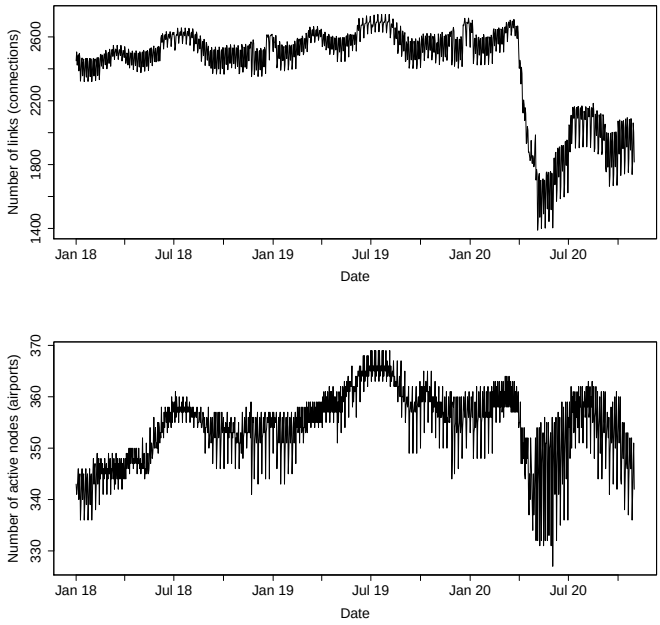
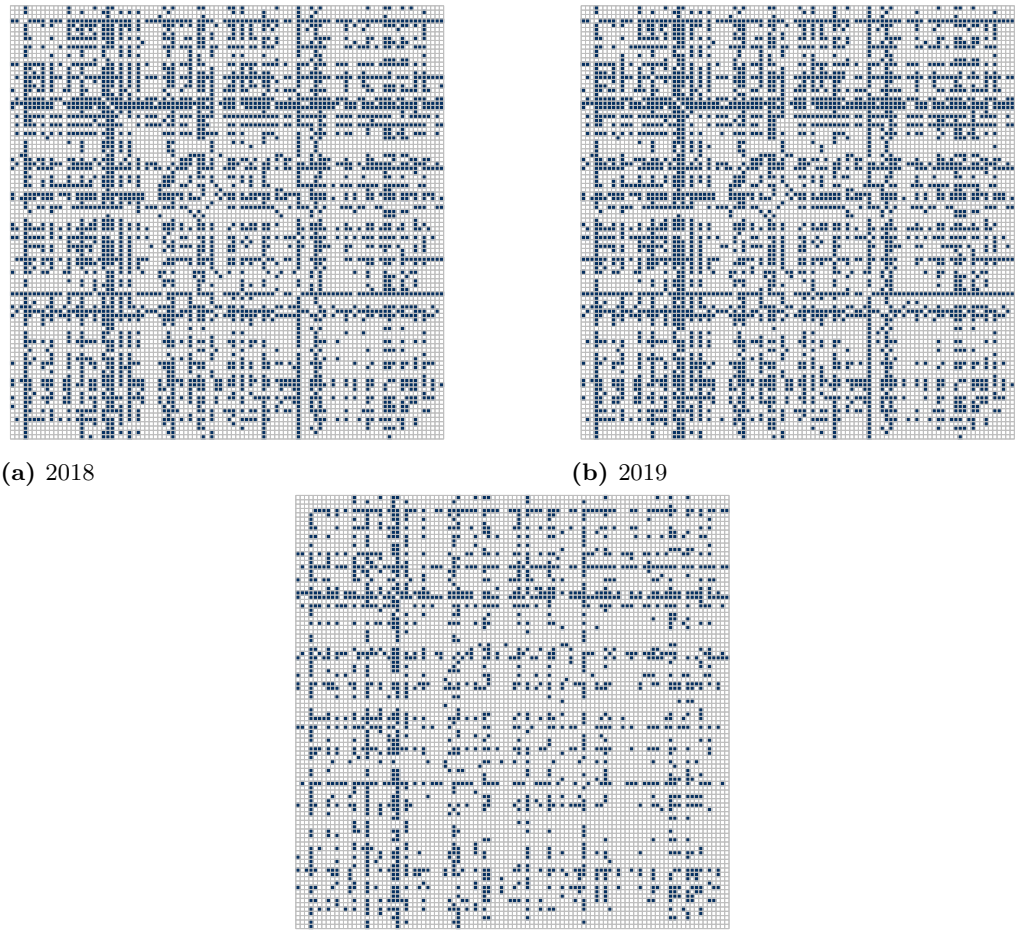


Figure 6: Link and node amount for the observed period of time

This can be underlined numerically as well. Regarding the first Tuesday in May, the percentage of the connections which include the five most central airports is 15.06% in 2018, 15.36% in 2019 and 21.50% in 2020.

Contrary to the simulation study, in which the networks were generated independently, it is obvious that the data is somewhat autocorrelated. It is likely that a connection is active if it was active the day or week before - especially regarding central airports. Consequently, we treat the emerging series of metric values as dependent over time. There are several ways to handle this issue when creating control charts in practice. See Knoth and Schmid, 2004 or Montgomery, 2007 for a variety of approaches. We proceed as follows: for each series we fit models of the ARIMA-class and use the AIC to individually decide which parameter setup for the ARIMA-models achieves the best fit. Subsequently, we apply residual charts which are based on the residuals of the corresponding ARIMA-model. We expect a bit more variation in the real-world data than in the simulation study - thus, we aim to identify rather large changes and, therefore, utilize Shewhart charts for individuals instead of EWMA charts. As convenient in this context [Montgomery, 2007], the residuals are assumed to follow a normal distribution with mean 0 and the control limits are set according to three standard deviations (3σ-rule).

We use the full year 2018 and the first half of 2019 for Phase I, which are suitable to represent the in-control state, as there were no major influences on the aviation business in this period. Consequently, Phase I and Phase II consist of 548 and 487 observations, respectively. We would like to generally point out that in some practical applications small changes from the



(a) 2018 **(b)** 2019

(c) 2020

Figure 7: Adjacency matrices of the 100 busiest airports for the first Tuesday in May. Blue dots signalize a connection, white dots otherwise.

in-control state in Phase I are possible. However, if they only occur very sporadically like here (e.g. the sole signal in Phase I of Figure 8), these time points may be retained. If major deviations are observed, we suggest to analyze and eliminate the cause and remove the points from the procedure.

For the analysis of the performances, see Figure 8 for the control chart of the spectral norm which can be seen as representative for the charts of the other metrics. For some metrics we observe occasional alarms in summer 2019 or winter 2019 which are probably due to the increased flight traffic during holiday time, but the anomalous behaviour since spring 2020 is strongly detected by each of them. The fact that all metrics work reasonable may seem somewhat surprising but is hugely down to the variety of different changes triggered by the pandemic as explained above.

The ability of the Frobenius norm to detect the anomalies is due to the heavily decreased link amount, while this is also the main reason why the upper and lower limit (see Section 2.4.1) of

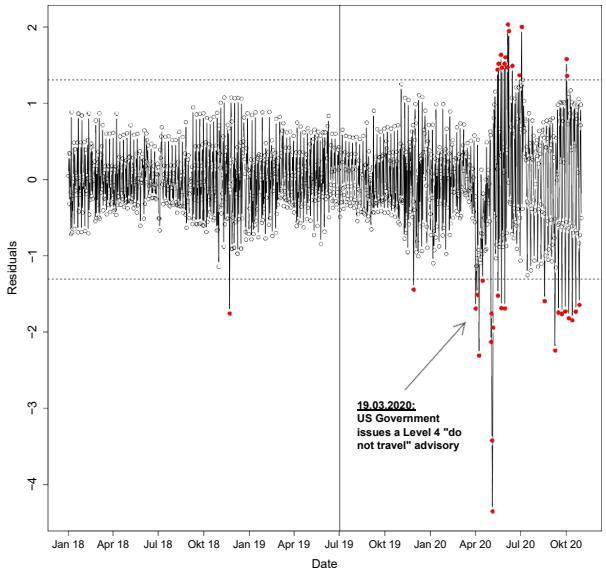


Figure 8: Shewhart Individuals Residual Control Chart for the spectral norm after fitting an ARIMA(2,1,0)-model. Phase I ends in July 2019. Red dots signalize alarms.

the spectral norm both tend to decrease - which affects that the spectral norm decreases as well.

Regarding the average centralities, we have to consider the results of Section 2.4.2. Due to the heavily decreased density, the scores $c(v)$ change noticeably regardless of which centrality criterion is used. However, due to the activity of fewer nodes, the factor $1/n$ gets larger and therefore weakens the effect for those criteria for which the values of $c(v)$ decrease (closeness, degree and eigenvector). Nevertheless, it is obvious from Figure 6 that the effect on the link amount is larger than on the node amount and, consequently, these metrics are handily able to detect the change. Similar reasons apply for the behaviour of the average path length.

The change in the network structure towards a greater centralization means that the centrality deviations can detect the anomalies as well. As explained in Section 2.4.2, the centrality scores $c(v)$ change due to the more sparse network but the most central node is not affected as much, because the central airports were able to conserve most of their flight routes.

The transitivity benefits from a less clustered pattern during the pandemic. This means that most destinations are reachable from each other but not directly anymore, because only the basic mobility was maintained. Consequently, the percentage of triangles, i.e. three airports for which a direct connection between each of them is active, clearly reduces.

We would like to note that a similar data set was analyzed with the usage of a model-based approach in Malinovskaya and Otto, 2021 which was also reliably able to identify the changed network pattern during the pandemic. However, a direct comparison shows the advantages of

Table 9: Performances of the metrics in various situations.

Change-Type	F	S	C_m	C_d	D_m	D_d	B_m	B_d	E_m	Tr	AvP
GLC	++	++	+	−	++	−	+	−	o	o	+
LLC	+	+	o	+	o	+	o	+	++	o	o
GNC	−−	++	+	−	+	−	+	−	o	o	+
LNC	−−	o	o	+	−	+	o	+	++	o	o

++ = suitable, + = mostly suitable, o = moderate performance/dependent on other circumstances, − = rather not suitable, −− = not suitable

the metric-based approach in such relatively simple scenarios. Firstly, this includes flexibility as the whole US air traffic was considered while a model-based-approach is usually limited to a fixed node set (or at least to low node dynamics) which usually leads to a failed usage of all available information. Secondly, the interpretations of the change-points are more understandable, since we were easily able to give explanations why each metric signaled a change. By using a more complex model-based approach on the other hand, it is usually not obvious why some changes happened.

2.7 Conclusion

Detecting Changes in dynamic networks is a challenging task. Due to the complex structure of network data, a variety of different types of changes are possible. In this paper, we proposed a general categorization of possible changes which is based on the principle that every structural element of networks can trigger a change. We distinguished between global changes, which affect the whole network equally, and local changes which only affect parts of the networks. Subsequently, we showed how popular network characteristics behave in those scenarios and analyzed their suitability of detecting them. The results can be used for change point detection, online monitoring, testing procedures and further network analyses. See Table 9 for a qualitative assessment of the performances of the considered metrics based on the presented results and explanations in this work.

We underlined our interpretations with a simulation study, where we monitored the different situations with EWMA charts. Subsequently, we showed the application in a real world example, where we monitored the US Flight network in a time-dependent setup with the help of ARIMA-models and Shewhart charts.

Note that the usage of network metrics in terms of monitoring is a straightforward approach which is especially effective when dealing with flexible network structures. In particular, this is the case if the user has little knowledge about the network structure and behaviour, or if the node set is not fixed or even is anonymous. Furthermore, we like to note that the performance might be increased in many situations, if a few metrics are monitored simultaneously in a

multivariate setup (e.g. MEWMA). A possible set of metrics can then individually be chosen with the help of the results presented throughout this paper.

3

Online monitoring of dynamic networks using flexible multivariate control charts

Based on: Flossdorf, Jonathan, Roland Fried & Carsten Jentsch (2023): "Online monitoring of dynamic networks using flexible multivariate control charts", In: *Social Network Analysis and Mining* 13, Article No. 87, DOI:10.1007/s13278-023-01091-y.

Abstract

Change-point detection in dynamic networks is a challenging task which is particularly due to the complex nature of temporal graphs. Existing approaches are based on the extraction of a network's information by the reduction to a model or to a single metric. Whereas the former one requires restrictive assumptions and has limited applicability for real-world social networks, the latter one may suffer from a huge information loss. We demonstrate that an extension to a well-balanced multivariate approach that uses multiple metrics jointly to cover the relevant network information can overcome both issues, since it is applicable to arbitrary network shapes and promises to strongly mitigate the information loss. In this context, we give guidelines on the crucial questions of how to properly choose a suitable multivariate metric set together with the choice of a meaningful parametric or non-parametric control chart and show that an improper application may easily lead to unsatisfying results. Furthermore, we identify a solution that achieves reasonable performances in flexible circumstances in order to give a reliably applicable approach for various types of social networks and application fields. Our findings are supported by the use of extensive simulation studies and its applicability is demonstrated on two real-world data sets from economics and social sciences.

3.1 Introduction

Dynamic networks play an important role in many different application fields nowadays, ranging from biological [Bassett and Sporns, 2017; Prill et al., 2005] and social sciences [Carrington et al., 2005; Sarkar and Moore, 2005] to logistic and transportation processes [Lee and Dong, 2009]. Suppose we observe a dynamic network $\mathcal{D} = \{D_t, t = 1, \ldots, T\}$ which is a sequence of snapshots of the network of interest at various time points t. Each of these single networks D_t consist of a set of nodes V_t that may be connected through a set of links E_t. Note that we do not only allow the number and positions of edges to differ between different time points, but also the number of nodes may change. It is often of interest to decide whether there are differences in the (dynamic) stochastic network generating process between different time points, e.g. due to a changed consumer behavior in marketing networks, an increased communication in social networks, or a failure of a working machine in a manufacturing process. Other scenarios involve financial market analysis [Durante and Dunson, 2014], network traffic monitoring [Sun et al., 2006], or connectomic applications [Durante et al., 2017]. Relevant statistical analysis procedures for such tasks are two-sample tests and change point detection. Although our results can be employed for classical testing procedures as well, we focus on the latter one in our work.

The main purpose of change point detection [Basseville and Nikiforov, 1993; Montgomery, 2012] is to identify time points at which the structure of a dynamic network changes in a meaningful way. Traditionally, there are two perspectives towards this issue. One approach is to observe the whole sample $\mathcal{D} = \{D_1, \ldots, D_T\}$ of interest first and to decide afterwards if one or more changes have happened (offline change point detection). Another approach is to monitor the process sequentially in real-time in order to make an immediate decision at each newly observed time point (online monitoring). Our results hold for both scenarios, but for clarity of exposition we mainly focus on the usually more difficult and, in practice, often more relevant task of online monitoring in the following.

In classical data setups of statistical process control, a time series of scalar values is monitored which means that the detection is mostly limited to a shift of the mean or variance of the observed measurement. Statistical network data, however, covers a lot more information compared to classical data setups (relationships, intensities etc.). Dynamic networks even add another dimension by the consideration of the time component. This has, on the one hand, the advantage of offering a more informative representation of an underlying system enabling e.g. a better interpretation of observed changes. On the other hand, the derivation of statistical inference approaches gets more challenging. For the task of online monitoring, there are two main issues in this context. First, it is not clear how a change might look like and, second, a direct transfer of traditional monitoring approaches to dynamic network data is hardly feasible as the process is described by a time series of networks and not by a time series of scalar values

anymore. It is therefore crucial to a) be aware of possible changes in network data and b) construct monitoring strategies that are suitable for network data.

3.1.1 Related Work

Regarding a), it is not intuitively clear how a change may look like, since there is not only one but many, partly dependent, components which may trigger a change in network structure. A straightforward definition is presented in Ranshous et al., 2015 by assigning a change to time point t, if $|f(D_t) - f(D_{t-1})| > c_0$ and $|f(D_t) - f(D_{t+1})| \leq c_0$ for some scoring function $f : D_t \mapsto \mathbb{R}$ and a threshold c_0. However, this approach is largely limited to the suitability of the applied scoring function as it addresses only those changes, for which $f(.)$ is able to capture the relevant information. In a more comprehensive context, the need of categorizing network changes with respect to their structural levels including nodes, communities or subgraphs is mentioned in Hewapathirana, 2019. Such a general categorization to handle this issue is presented in Flossdorf and Jentsch, 2021. It covers global as well as local changes of single components (e.g. nodes and links) and addresses their combinations such that more complex structural changes (e.g. in blockmodels) are also considered.

Regarding b), monitoring strategies for dynamic network data are typically based on a reduction of the complexity of the underlying time series of networks. There exist various methods to do so which can be subdivided in model-based, embedding-based and metric-based approaches. For the model-based methods, a dynamic network model is fitted and the specific parameters or residuals are then monitored with a traditional control chart. Examples are state space models [Zou and Li, 2017], degree-corrected stochastic blockmodels [Wilson et al., 2019], temporal exponential random graph models [Malinovskaya and Otto, 2021], or Poisson regression models [Farahani et al., 2017; Motalebi et al., 2021]. However, the model-based approach commonly requires potentially restrictive assumptions like a fixed node set (same nodes for each time point, no node dynamics) or knowledge of the underlying network structure. These assumptions are often too strong, since they can only be used for a small field of applications. Furthermore, they allow to detect only a limited number of changes while ignoring those which does not affect the fitted model.

A related approach is the usage of embedding techniques. In a nutshell, the main goal of network embedding is to learn a mapping function in order to map each single node to a lower-dimensional vector. This results in a latent lower-dimensional feature representation of a graph that reduces noise and redudant information, but still aims to maintain important structural information [Cui et al., 2018]. There exist various of these network representation methods such as node2vec [Grover and Leskovec, 2016] and DeepWalk [Perozzi et al., 2014] that are developed based on random walks. Other approaches use matrix factorization [Belkin and Niyogi, 2001;

Ou et al., 2016]. Whereas these are specifically designed for static networks, there also exist versions for dynamic networks like temporalnode2vec [Haddad et al., 2020] and DynamicTriad [Zhou et al., 2018]. A thorough analysis of the stability of such embedding approaches can be found in Gürsoy et al., 2021, where also the importance for ensuring alignment is stressed as misaligned embeddings might negatively impact the performance for dynamic network inference tasks such as e.g. change detection. For the usage of change detection, embedding methods are particularly used to detect vertex-based changes in a time series of graphs like, e.g., detecting vandal users in online social networks using a vector autoregression approach [Li et al., 2019]. Moreover, Hewapathirana et al., 2020 propose a spectral embedding method that particularly addresses sparsity and degree heterogeneity. Lin et al., 2022 construct an embedding change detection approach with a focus on node coordinates in a latent space that are used to model edge dependencies. Further approaches involve Sun and Liu, 2018, Grattarola et al., 2019, Duan et al., 2020, and Xie et al., 2023.

Lastly, metric-based approaches summarize the network information by assigning a single metric or a combination of different metrics to each D_t. Hence, they are more flexible as they can be applied to arbitrary types of networks. Exceptions are similarity measures like DeltaCon [Koutra et al., 2016] or Graph Edit Distance [Bunke et al., 2007] which sequentially compare each D_t to a reference network. For those approaches a fixed node set is obviously required. This is not the case for any other network metric that is calculated with the sole information of D_t. Recent works used centrality metrics [McCulloh and Carley, 2011], matrix characteristics [Barnett and Onnela, 2016; Hazrati-Marangaloo and Noorossana, 2021], scan statistics [Neil et al., 2013], or further network metrics like the clustering coefficient [Kendrick et al., 2018]. The application under the consideration of time dependency is discussed in Ofori-Boateng et al., 2021. Obviously though, univariate metric-based methods tend to lead to flexibility issues. A single metric is only able to capture some, but not all information that might be relevant for change detection. Therefore, in Flossdorf and Jentsch, 2021, such metrics are analyzed in dependent and independent setups and evaluated with respect to their individual suitability in various situations. A multivariate usage of three network centrality metrics and the network density in a multivariate EWMA chart is studied in Salmasnia et al., 2020. However, a fixed node set as well as a Gaussian distribution for each involved metric is assumed which noticeably reduces the flexibility advantage that metric-based approaches have compared to the other mentioned categories of network change detection.

3.1.2 Contribution

We expand on the univariate results of the metric-based approaches and study a monitoring method that uses a multivariate set of metrics. We illustrate that such a multivariate approach overcomes the mentioned issues of model-based and other metric-based methods, since it is

ad-hoc applicable without restrictive assumptions and provides flexibility gains as the relevant information is now simultaneously captured by multiple metrics of various types. We base our analysis on theoretical considerations, simulation-based evidences and practical applications. The need of such a multivariate analysis is mentioned in the univariate literature [Flossdorf and Jentsch, 2021; McCulloh and Carley, 2011]. Contrary to the multivariate investigations in Salmasnia et al., 2020, we do not limit our analysis on a certain set of metrics or control chart and also allow for arbitrary network structures including a dynamic node set and a non-parametric setup. Our primary objective is to present a solution on how exactly a general multivariate procedure shall be implemented and performed. In this context, we identify three challenges that are crucial for a successful application: a) a sound choice of a set of network metrics, b) their combination with a suitable choice of control charting procedures, and c) the final interpretation of the results. We introduce strategical guidelines to solve these challenges and demonstrate that ignoring them could easily lead to erroneous conclusions and unsatisfying results. We identify a balanced solution that achieves reliable performances in various situations as well as propose solutions that are specialized for more specific scenarios. To support the flexibility of this study, we study both distribution-free and parametric monitoring schemes.

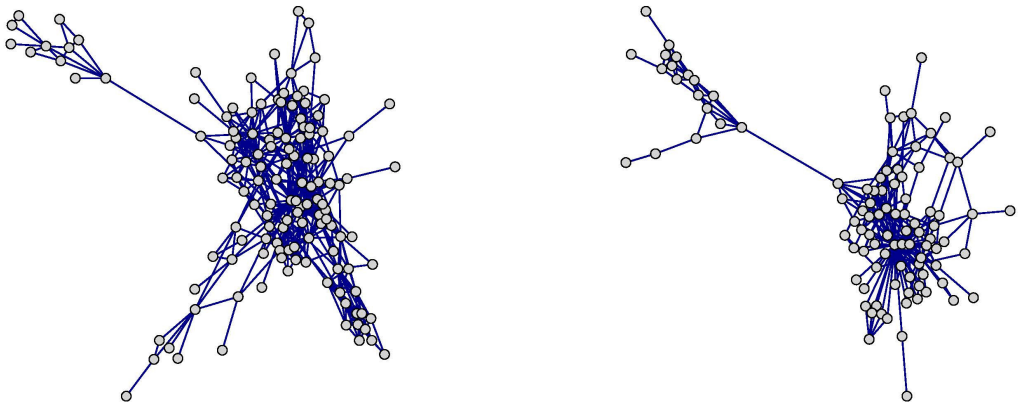
The paper is organized as follows: Section 3.2 offers a short recap of the necessary theory regarding change detection in network data and multivariate control charts. In Section 3.3, we formulate the multivariate network change detection procedure in details and study the suitability of different metric sets, various control charts, and the interaction of both. The results are supported by an extensive simulation study in Section 3.4 which confirms the reliability in various change situations and shows superior performance compared to univariate approaches. In Section 3.5, the procedure is applied to two real-world social network data sets. Section 3.6 contains some concluding remarks.

3.2 Existing Foundations

In this section, we recap the existing concepts of change detection in network data as well as of traditional statistical process control that are necessary for the introduction of the methodology in Section 3.

3.2.1 Changes in Network Data

The first challenge is to understand which types of changes may happen in network data. Because a dynamic network consists of various structural elements, a simple shift of location or scale parameters like in traditional scenarios does not exist. As explained above, we focus on a



(a) network before change **(b)** network after change

Figure 9: Real-world example for a mixed-type change (MTC): Enron E-mail communication.

flexible setup and prefer general types of changes [Hewapathirana, 2019] rather than specialized changes and therefore follow the change definition of Flossdorf and Jentsch, 2021. In this context, the idea is to consider the influence of each structural network element to obtain a thorough categorization of possible changes. These elements are a) links, b) nodes, and c) extra information that may be put on either nodes or links (i.e. covariates, attributes). Each element is assumed to be able to trigger a change either in a global or local manner. A short summary of all scenarios is listed below. Note that we do not consider changes caused by covariates here, since their type of occurrence is hugely dependent on the underlying application field.

- **Global Link Change (GLC):** The change is triggered by a significantly increased or decreased link amount. This is assumed to happen globally, i.e. the changed link probability affects each node equally.
- **Local Link Change (LLC):** Similar to GLCs, but the changed link behavior only affects a few nodes which either get more or less influence, i.e. the network structure changes to a more centralized or flat hierarchy.
- **Global Node Change (GNC):** The node amount increases or decreases significantly, because new nodes enter the network or existing ones leave it.
- **Local Node Change (LNC):** Only a few influential nodes enter or leave the network which results in a significant impact on the network structure.

While all of these changes may occur individually, it is likely that some of them happen simultaneously, e.g. a global increase of links may be the consequence of the entry of new

Table 10: Examples and Application Scenarios

Type	Examples
GLC	increased/decreased communication in communication networks changed activity in cyber networks (e.g. due to malware)
LLC	formation of new hotspots in disease networks changed route layout in transportation networks
GNC	new advertising strategy in customer networks addition of new destinations in tourism networks
LNC	restructuring of supply chains in logistics networks creation of new leading positions in profession networks

nodes in the network. This is referred to as mixed-type changes (MTC), which also enables capturing more complex change scenarios like those in blockmodels or subgraphs. A real-world social network example for MTCs is depicted in Figure 9. It shows the e-mail communication between employees of the former US company Enron. It is quite obvious that the overall structure of the network has not changed drastically between the both observed time points. Hence, we have no hint to local changes. However, we can observe a clearly reduced node and link amount in Figure 9b compared to Figure 9a. The network hence got more sparse both in terms of nodes and links which hints to a mixed-type change of GLC and GNC. Further typical examples and relevant application scenarios for the different types of change are given in Table 10.

3.2.2 Metric-based Network Monitoring

We already briefly discussed possible complexity reduction procedures in order to monitor a temporal series of graphs which involves model-based and metric-based approaches. The former one is quite restrictive and not applicable ad hoc, e.g. parametric assumptions have to be met. This also affects that only a few model-specific change types can be detected, e.g. LNCs and GNCs are ignored due to the common assumption of a fixed node set. Those restrictions are especially unfavorable, if the structure and behavior of the network of interest is not explicitly known beforehand. Dynamic networks are commonly quite prone to this issue due to their high dimensionality and potentially high dynamics.

The complexity reduction step of the metric-based procedure is even more radical, but those approaches provide a more flexible monitoring tool without restrictions for a broad application field. Consider that each network D_t of the dynamic network $\mathcal{D}$ is reduced to a scalar $f(D_t) = s_t$. Thus, the vector $\mathbf{s} = (s_1, \ldots, s_T)$ contains the captured information of the applied metric to $\mathcal{D}$. Common choices for global summary metrics are matrix norms, for instance:

- **Frobenius norm (F):**
  $\sqrt{\sum_{i=1}^{n_t} \sum_{j=1}^{n_t} a_{t,ij}^2}$.

- **Spectral norm (S):**
  $\max_{\|x\|_2=1} \|A_t x\|_2$.

These matrix norms are calculated based on the temporal series of adjacency matrices $A_t \in \mathbb{R}^{n_t \times n_t}$ of the networks D_t, where n_t is the number of nodes at time point t. For the sake of simplicity, we mainly focus on undirected and unweighted networks which means that a matrix entry $a_{t,ij} = 1$, if a link between nodes i and j exists, and $a_{t,ij} = 0$ otherwise. Regarding online monitoring, the Frobenius norm is studied in Barnett and Onnela, 2016, the spectral norm in Chen et al., 2021 and as it is equal to the largest eigenvalue for undirected networks also in parts in Hazrati-Marangaloo and Noorossana, 2021.

Further popular network metrics for change detection are centrality measures as studied in McCulloh and Carley, 2011, Salmasnia et al., 2020, or Ofori-Boateng et al., 2021. This includes

- **Closeness centrality (C):**
  $c_i^C = (\frac{1}{n-1} \sum_{j \in V} \mathrm{d}(i,j))^{-1}$.

- **Degree centrality (D):**
  $c_i^D = \sum_{j=1}^{n_t} a_{t,ij}$.

- **Betweenness centrality (B):**
  $c_i^B = \sum_{i \neq j \neq l} \frac{\sigma_{jl}(i)}{\sigma_{jl}}$.

- **Eigenvector centrality (E):**
  $c_i^E = \frac{1}{\lambda} \sum_{j=1}^{n_t} a_{t,ij} c_j^E$.

In this context, $d(i,j)$ denotes the shortest path between two nodes i and j, $\sigma_{jl}(i)$ the number of shortest path between two nodes j and l (that pass through node i), and λ the largest eigenvalue of A_t. Whereas the matrix norms are global metrics for the whole network D_t, the centrality scores are locally defined for each node. To transform them into a global network metric, we consider two approaches. On the one hand, this involves the average score over all nodes, i.e.

$$c_{\text{avg}} = \frac{1}{n} \sum_{i \in V} c_i.$$

On the other hand, we can use a scale metric by taking the deviation to the largest observed score

$$c_{\text{dev}} = \sum_{i \in V} (\max_{j \in V}(c_j) - c_i).$$

We specify the used version in the following by noting an m or d index (e.g. C_m for mean Closeness and C_d for Closeness deviation). Note that for the eigenvector centrality, both versions are affine linear transformations to each other and therefore achieve equal monitoring performances. It is possible to use any other network metric as well, e.g. the Average Path Length or the Clustering Coefficent like in Kendrick et al., 2018, but many are strongly related to the presented ones and do not contribute added value [Flossdorf and Jentsch, 2021].

3.2.3 General Online Monitoring Procedure

The main goal of online monitoring is to detect anomalies in a process as soon as possible after their occurrence. Typically, the process of interest is subdivided into two phases. In Phase I, it is assumed that the process is somewhat stable and reliably represents the typical state of the underlying system without meaningful deviations. The system is then called to be *in-control*. In Phase II, the actual monitoring takes place by deciding if an incoming signal sufficiently matches the in-control state. An alarm is triggered if this is not the case and the signal is classified as *out-of-control* [Basseville and Nikiforov, 1993].

Consider $\{Y_t, t = 1, \ldots, T\}$ to be a sequence of a random variable of interest with conditional density $P_\theta(Y_t \mid Y_{t-1}, ..., Y_1)$ and τ to be the unknown change time. If $\tau > t$, then the conditional density parameter θ is constant with $\theta = \theta_1$. For $\tau \leq t$, it applies $\theta = \theta_2$. The goal is to detect the anomaly as soon as possible with a fixed rate of false alarms before τ. An estimation of θ_1 and θ_2 is often not necessary, but might be useful for interpretation purposes regarding possible reasons for a change.

In terms of the practical usage, control charts are applied. First, a metric x_t is chosen which a) can be calculated for each time point t, b) covers and represents most relevant aspects of the behavior of the system (i.e. Y_t), and c) is able to identify all considered changes by a sensitive reaction to them. This metric serves as the main input for the control statistic z_t which is calculated for the process at each time point t. Depending on the setup, z_t can be the metric itself (e.g. in memory-free Shewhart charts) or some sort of transformation $z_t = g(x_t)$ (e.g. in memory-based EWMA or CUSUM charts). In Phase I, z_t is expected to represent the in-control state of the system and, hence, to be a stable process without meaningful deviations. This information is used to define the upper and lower control limits h_u and h_l. Those limits are chosen under the consideration of the desired rate of false alarms and can be derived by parametric or distribution-free procedures. In Phase II, if we observe $h_l \leq z_t \leq h_u$, the process is deemed in-control. Otherwise an alarm is triggered at time point t to signal a detected change.

3.2.4 Control Charts for Traditional Multivariate Data

In practice, it is customary to improve the monitoring procedure by taking more process metrics into account. Their independent usage in a univariate manner is possible but not recommended, since it is inefficient and may result in erroneous conclusions [Montgomery, 2012]. Hence, the construction of multivariate control charts, which consider the metrics jointly, is of interest.

In this context, the most basic multivariate chart is the Hotelling T^2 chart. Suppose we observe a vector $\mathbf{x_t} = (x_{t1}, ..., x_{tp})$ of p different process metrics at each time point t, then the corresponding control statistic is calculated by

$$z_t = (\mathbf{x_t} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x_t} - \bar{\mathbf{x}}),$$

where $\bar{\mathbf{x}}$ and $\mathbf{S}$ are the sample mean vector and covariance matrix of the underlying observations. Since z_t mainly takes the squared deviation to the sample mean into account, it is non-negative and we expect values near zero if the process is in-control. Therefore, only an upper control limit has to be derived. This can be done under parametric assumptions with an approximation via the F-distribution which yields

$$h_u = \frac{p(n+1)(n-1)}{n^2 - np} F_{\alpha,p,n-p},$$

where n is the number of observations in Phase I and α the desired false alarm rate.

In many practical applications, the met distributional assumption might not be justified which can have a strong negative impact on the monitoring performance. To avoid such issues, the usage of non-parametric techniques seems promising. In this context, a bootstrap approach was proposed in Phaladiganon et al., 2011, which is able to efficiently handle the monitoring process, even if the observed data is non-Gaussian or unknown. It works as follows. First, the statistic z_t is calculated for all T observations of Phase I as before, which yields the vector $\mathbf{z} = (z_1, ..., z_T)$. Subsequently, B Bootstrap samples with sample size T are drawn from $\mathbf{z}$ and for each of those samples the $(1-\alpha)$-quantile is calculated. The upper control limit h_u is then determined by taking the average over those values.

The Hotelling T^2 charts are multivariate extensions of a univariate Shewhart-type control chart, because they only use information of the current observation which makes them rather insensitive to small shifts. Memory-based control charts like exponential weighted moving average charts (EWMA) overcome this issue, the control statistic of a multivariate version (MEWMA) [Lowry et al., 1992] is defined as

$$z_t = \mathbf{m}_t' \mathbf{S_t}^{-1} \mathbf{m}_t,$$

where $\mathbf{m_t}$ is recursively defined as

$$\mathbf{m}_t = \lambda(\mathbf{x}_t - \bar{\mathbf{x}}) + (1 - \lambda)\mathbf{m}_{t-1}.$$

The estimation of the covariance matrix is given by

$$\mathbf{S}_t = \frac{\lambda}{2 - \lambda}[1 - (1 - \lambda)^{2t}]\mathbf{S},$$

where $\mathbf{S}$ is the estimated covariance matrix given all observations from Phase I. While the formula of the control statistic is quite similar to the Hotelling T^2 chart, the main difference lies in the intermediate step of calculating $\mathbf{m}_t$, where the smoothing parameter $\lambda \in [0, 1]$ serves as a factor for providing weights to past observations and the current one. For $\lambda = 1$, the MEWMA setup corresponds to the Hotelling T^2 chart. Optimal control limits depending on λ, the number of variables p and the desired false alarm rate can be found in several works [Knoth, 2017; Prabhu and Runger, 1997]. In general, MEWMA charts with small values of λ are rather robust to the normal assumption yielding satisfying results for different distributions of the underlying data [Montgomery, 2012].

3.3 Multivariate metric-based monitoring solutions

After the recap of the theoretical background of the existing foundations in Section 3.2, we now move on to the practical implementation of a monitoring setup for network data by combining and adapting multivariate control chart schemes with an intelligent choice of a set of network metrics. Making use of network metrics for the monitoring of networks, first, the practitioner has to choose the (multivariate) set of metrics in conjunction with a suitable control chart procedure. While there is a whole variety of (and even more combinations of) networks metrics that could be used, it is generally unclear which combination of such network metrics should be used in which dynamic network scenario to detect deviations from the control state as reliable as possible. As the further steps are rather straightfoward, that is, the calculation of the corresponding control statistic and control limits (Phase I), the monitoring of new observations (Phase II) and how to stop at a detected change, the interpretation of the change is also not straightforward. In this context, note that this work aims to present a general procedure and particularly focuses on developing an ad-hoc solution that is applicable to arbitrary network shapes and application fields and enables a reliable change analysis in various situations. Therefore, we present general guidelines for different categories of application scenarios and identify the most flexible solution that performs reliably even in unknown circumstances.

Table 11: Performances of the metrics in various situations.

Change Type	F	S	C_m	C_d	D_m	D_d	B_m	B_d	E
GLC	++	++	+	−	++	−	+	−	o
LLC	+	o	−	+	o	+	o	+	++
GNC	−−	++	+	−	+	−	+	−	o
LNC	−−	−	o	+	−	+	−	+	++

++ = suitable, + = mostly suitable, o = moderate performance/dependent on other circumstances, − = rather not suitable, −− = not suitable

3.3.1 Selection of a suitable set of metrics

We use the metrics that were presented in Section 3.2.2 as these are the common choices for extracting the information of a network [Barnett and Onnela, 2016; Chen et al., 2021; Hazrati-Marangaloo and Noorossana, 2021; McCulloh and Carley, 2011; Ofori-Boateng et al., 2021; Salmasnia et al., 2020]. Because the information of a network gets reduced to a single scalar value, the resulting information loss is typically quite large. It is therefore crucial to be aware of the information a metric is able to capture in order to understand which type of change it is able to detect which was discussed in Flossdorf and Jentsch, 2021. See Table 11 for a short summary of the individual suitability of the considered metrics in the presented change scenarios.

Due to the described information loss, it is not too surprising that no single metric is able to perform well in all scenarios. However, for each type of change, there exist multiple metrics that work reasonably well. Hence, it is a natural approach to use multiple metrics jointly in order to capture various pieces of information to mitigate the loss. Formally, for each D_t, a vector $\mathbf{s}_t = (s_{t1}, ..., s_{tp})$ of p different scores is calculated at each time point t. In this work, we mainly focus on sets with $p = 3$.

In this context, the main challenge is the choice of a suitable set of metrics. The main statement of Table 11 is that most metrics either perform reasonably well in change situations that affect either the network globally (i.e. GLC, GNC) or in local change scenarios (i.e. LLC, LNC). Based on this, we may classify most metrics into two different performance groups A = {S, C_m, D_m, B_m}, which perform well in global change scenarios, and group B = {C_d, D_d, D_b} that perform superior in local setups. Remaining are the Frobenius norm, which can handle link changes but ignores node changes, and average Eigenvector Centrality that is theoretically affected by all change types but sometimes to a lesser extent.

The final choice of a suitable set of metrics is dependent on the goal and the expectations of the application. Based on the described univariate behaviour of the presented metrics, we propose to consider the following multivariate monitoring strategies.

- **I - Balanced Setups**: This category represents the most flexible monitoring strategy. The goal is to achieve a reliable performance for as many as possible change types. Particularly, changes of moderate to high intensity should be detected. Corresponding sets include one metric from each of the performance groups A and B. The third metric is the average eigenvector centrality. This balanced setup is a promising candidate for ad-hoc applications in which users do not know what to expect and how a change might look like (e.g. networks with high dynamics like social networks). Example setup: SB_dE.

- **IIa - Balanced Setups with a focus on global changes**: In this category it is still of interest to be sensitive to as many change types as possible, but with a clear focus on the detection of global changes. This category involves all 2 vs. 1 combinations (i.e. two metrics out of group A and one out of group B). Example setup: SD_mC_d.

- **IIb - Balanced Setups with a focus on local changes**: The same as IIa, just with a major focus on local changes (i.e. one metric out of group A and two out of group B). Example setup: B_dD_dS.

- **IIIa - Unbalanced setups for global changes**: It is only of interest to detect global changes in the network - even those which are characterized by a small change intensity. The detection of other change types is not relevant. The corresponding metric sets are constructed with three metrics out of group A. Example setup: C_mD_mS.

- **IIIb - Unbalanced setups for local changes**: The same as IIIa, just for local changes. The metric sets are constructed with three metrics out of group B. Example setup:$C_dD_dB_d$.

- **IV - Setups with a particular focus on link changes**: This category represents all metric sets in which the Frobenius norm is part of, since this metric is specialized to detect link changes. Changes purely triggered by nodes can also be emphasized by combining the Frobenius norm with other metrics that are sensitive to it. Example setup: FSB_d.

In a nutshell, the idea behind Category I is to use one metric out of both classes A and B, in order to capture various types of information. The usage of the average eigenvector then provides some neutral perspective. This *balanced* setup seems to be a promising candidate for an ad-hoc application in which users do not know what to expect and how a change might look like (e.g. networks with high dynamics like social networks). The other Categories offer some more *unbalanced* setups by 2 vs. 1 and 3 vs. 0 combinations. These are constructed for more specialized cases where the user might be interested in detecting some particular changes which frequently occur in the corresponding application (see examples in Table 10). Finally, Category VI emphasizes link changes more by taking the Frobenius norm into account.

As mentioned, we used $p = 3$ as the number of considered metrics for the development of the categorization scheme above. The usage of other values is of course possible. However, we would like to note that a higher value of p is in theory helpful for capturing more information, but is also prone to more uncertainty picked up by the monitoring procedure, which might result in power losses. Moreover, some of the metrics might be highly correlated due to a similar definition which would make the monitoring procedure less efficient for higher p and might even lead to erroneous conclusions. Moreover, values of $p > 3$ do likely not contribute added value as there are the two explained performance groups. Considering this, $p = 3$ seems to be a good trade-off between information capturing and maintaining flexibility. This is also validated by our simulation study in Section 4.

A particular metric set with a value of $p = 4$ is considered in Salmasnia et al., 2020. For the remainder of this paper, we denote this set by SAL. The included metrics are deviation metrics for the Betweenness, Degree, and Closeness centralities with a similar definition to B_d, D_d, and C_d. Additionally, the network density is used as a fourth metric. Overall, this metric set can be classified in our Category IIIb as it uses three metrics that are specifically sensitive to detect local changes. As we will see in the simulation study, the additional consideration of the density does not change much compared to the $B_dD_dC_d$ set as these metrics are highly correlated and quite dominant in their multivariate combination.

3.3.2 Selection of a suitable multivariate control chart

The choice of a suitable control chart setup is as important as the choice of a metric set. The monitoring performance of the parametric Hotelling T^2 chart is dependent on the quality of the fit of the applied F-distribution. To the best of our knowledge, no complete asymptotic inference has yet been derived for the considered network metrics due to their complex nature. Consequently, putting parametric assumptions on their joint distribution seems rather implausible. The parametric Hotelling T^2 chart is known to react rather sensitively to violations of its distributional assumption [Stoumbos and Sullivan, 2002] and might suffer from reliability issues in this context. We expect that this is especially the case for rather unbalanced multivariate sets of metrics, since their marginal distributions tend to be more similar to each other and are sensitive to the same impact factors, which may result in a more skewed joint distribution. This behaviour can particularly be observed for sets of Categories IIIa and IIIb and also for the presented one of Salmasnia et al., 2020. For more balanced setups, this effect is likely to be weakened, because different sensitivities are involved. Furthermore, we expect a worse performance of the Hotelling T^2 chart for lower false alarm rates α, because the corresponding control limit h_u depends on the $(1-\alpha)$-quantile of the applied distributional assumption. Higher quantiles are likely to be bad approximations for the corresponding

quantiles of the empirical distribution, which is - for very low α - sensitive to the observed extreme values that might especially play a role for rather unstable and high-dynamic processes like networks. Overall, this effect tends to be less pronounced for larger values of α as the quantile of interest is shifted more towards the center of the distribution. The explained impacts affect the MEWMA chart as well, but to a lesser extent. Due to the smoothing of the control statistic that involves the consideration of past observations, the chart is noticeably more robust against non-Gaussian behavior. This is particularly the case for lower values of the smoothing parameter λ which weakens the individual influence of the current observation. However, note that inertia issues might be a consequence of this [Lowry et al., 1992].

Finally, the non-parametric Hotelling T^2 chart might be the safest choice for a reliably constructed control chart when using the considered metric sets. As the bootstrap procedure is directly dependent on the empirical distributions, we expect the chart to be more robust against various types of metric sets, i.e. to perform on a similar level for all sets. Obviously, its quality increases for larger sample sizes (i.e. longer in-control phase), and the bootstrap procedure ensures that it works reasonably in most cases of lower sample sizes as well. However, in the latter cases, its performance might not be superior to the parametric candidates anymore as we will see in Section 3.4.

3.3.3 Interpretation of the results and further analyses

To conclude, we recommend using a balanced metric set like $\mathrm{SB}_d\mathrm{E}$ together with the non-parametric Hotellings T^2 chart for the most reliable solution in flexible scenarios. Regarding the interpretation of the monitoring results, note that we monitor a temporal series of networks $\mathcal{D} = \{D_t, t = 1, \ldots, T\}$ instead of a simple process variable Y_t. Hence, the change parameter θ gets more complex (see Section 3.2.3) making its interpretation in a change situation all the more important. This especially concerns the purpose of maintaining transparency and reliability of the monitoring tool. In practice, we propose to stop after a detected change and to analyze the corresponding network D_t carefully. This can be handily done by descriptively analyzing the univariate values of the used metrics and applying their interpretations given in Table 11 in order to identify the underlying change type. Precise intepretation examples are given in Section 3.5.

While focussing on temporally independent setups and the related challenges in this paper, the presented procedure can be handily extended to allow also for time-dependency. For instance, this can be done by fitting an ARIMA model to the multivariate series of metrics in order to monitor its residuals similar to Flossdorf and Jentsch, 2021; Ofori-Boateng et al., 2021; Pincombe, 2005.

Table 12: Data generation processes using Erdös-Renyi (ER) graphs.

Type	Data Generation
GLC	In-Control: ER graph with probability p and n nodes Out of Control: ER graph with probability $\tilde{p}$ and n nodes
LLC	In-Control: ER graph with probability p and n nodes Out of Control: heterogenous ER graph with a changed probability $\tilde{p}$ for k central nodes
GNC	In-Control: ER graph with at each time point a randomly chosen node size in $[n - n \cdot d, n + n \cdot d]$ and a randomly chosen link amount in $[m - m \cdot d, m + m \cdot d]$. Out of Control: ER graph with at each time point a randomly chosen node size in $[\tilde{n} - \tilde{n} \cdot d, \tilde{n} + \tilde{n} \cdot d]$ and a randomly chosen link amount in $[m - m \cdot d, m + m \cdot d]$.
LNC	In-Control: ER graph with probability p and at each time point a random node size $n_{\text{flex}} \in [n - n \cdot d, n + n \cdot d]$. Out of Control: heterogenous ER graph with at each time point a random node size $\tilde{n}_{\text{flex}} \in [\tilde{n} - \tilde{n} \cdot d, \tilde{n} + \tilde{n} \cdot d]$ and changed probability $\tilde{p}$ for k central nodes and a probability of $\frac{pn_{\text{flex}} - \tilde{p}k}{\tilde{n}_{\text{flex}} - k}$ for all other nodes.

3.4 Simulation Study

To underline our findings, we execute an extensive simulation study in the following. We generate numerous example situations of each described change type and analyze the performances of the proposed multivariate metric strategies in combination with the described control chart procedures. Additionally, we compare their results with the corresponding univariate metric performances as studied in McCulloh and Carley, 2011, Ofori-Boateng et al., 2021, or Barnett and Onnela, 2016 and with the multivariate approach SAL of Salmasnia et al., 2020. Recall that a comparison with typical model-based and embedding-based approaches is not feasible here in a meaningful way, as our study is designed to detect flexible changes without imposing any high-level assumptions on the network structure (e.g. fixed node set, model assumptions). In a second part, we extend the study by analyzing more practical relevant mixed-type changes in different situations of stochastic blockmodels. The code for the simulation study can be found in a GitHub repository[1] created by the authors.

3.4.1 Setup

We generate sampling data for each of the four described change types of Section 3.2.1 (GLC, LLC, GNC, and LNC). For each of those, we execute three sub-situations with varying change

[1] github.com/jonathanFlossdorf/NetworkMonitoring

Table 13: Parameter setup for each situation according to Table 12 with $n = 50$.

Type	Intensity	Parameter Setup
GLC	small	$p = 0.4$, $\tilde{p} = 0.43$
	moderate	$p = 0.4$, $\tilde{p} = 0.45$
	heavy	$p = 0.4$, $\tilde{p} = 0.5$
LLC	small	$p = 0.35$, $\tilde{p} = 0.45$, $k = 5$
	moderate	$p = 0.4$, $\tilde{p} = 0.6$, $k = 2$
	heavy	$p = 0.35$, $\tilde{p} = 0.65$, $k = 1$
GNC	small	$d = 0.05$, $m = 800$, $\tilde{n} = 55$
	moderate	$d = 0.05$, $m = 800$, $\tilde{n} = 60$
	heavy	$d = 0.05$, $m = 800$, $\tilde{n} = 75$
LNC	small	$d = 0.15$, $p = 0.4$, $\tilde{n} = 54$, $\tilde{p} = 0.6$, $k = 4$
	moderate	$d = 0.15$, $p = 0.4$, $\tilde{n} = 52$, $\tilde{p} = 0.7$, $k = 2$
	heavy	$d = 0.15$, $p = 0.4$, $\tilde{n} = 51$, $\tilde{p} = 0.8$, $k = 1$

intensities (small, moderate, and heavy). This results in 12 scenarios in total, which are repeated 1000 times to obtain a reliable performance analysis. The data generating processes as well as their different parameter setups to control the intensity are given in Tables 12 and 13.

For each situation, we simulate dynamic networks of length $T = 1400$ and use the first 1000 observations as Phase I in order to reliably calibrate the control chart. The change time is set to happen at time point $\tau = 1050$. The control limits of the control charts are set with a false alarm rate of $\alpha = 1\%$. While the rather large number of observations in Phase I is interesting to obtain meaningful findings about the theoretical performances and the suitability of the metric sets, we are aware that, from a practical point of view, those numbers are usually hard to provide in real-world applications. To this purpose, we examine the performances in practically more relevant simulation settings in Section 3.4.4 and analyse real-world networks in Section 3.5.

For all scenarios, we evaluate the performances of all triple combinations of the presented metrics and particularly focus on the results of the multivariate metric set categories that were derived in Section 3.3.1. We evaluate their performance in combination with all three presented control chart procedures and compare the results with the univariate approaches. As performance measures, we use ARL_0 and ARL_1. The ARL_0 is defined as the in-control average run length which can reach an optimal value of $\frac{1}{\alpha} = 100$ in our setup. On the other hand, ARL_1 calculates the post-change average run length which measures the delay to detection, i.e. the number of time points an alarm is sent after the actual change has occurred.

To maintain comparability across the univariate and multivariate setups, we compare memory-free settings and memory-based procedures separately. Hence, the parametric and non-parametric Hotellings T^2 charts are compared with the Shewhart chart and the MEWMA

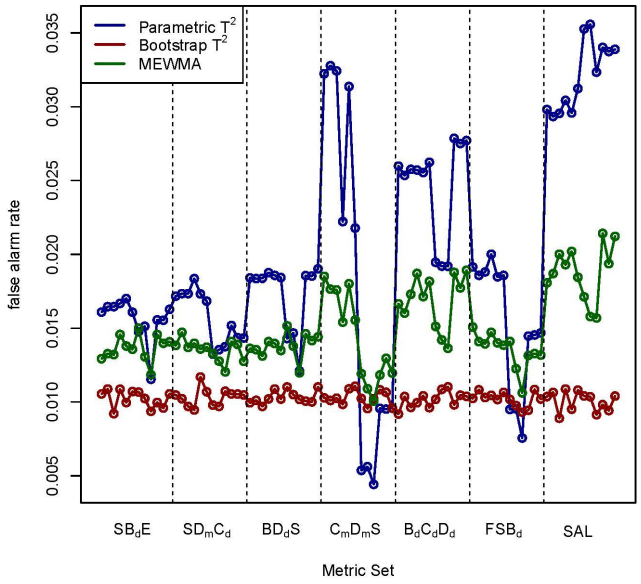


Figure 10: Average empirical false alarm rates for all simulated situations. The desired value for α is 0.01.

chart with the EWMA procedure. For the latter ones, we try different smoothing parameters $\lambda \in \{0.1, 0.2, ..., 0.7\}$ and report the ones with the best results.

3.4.2 Phase I Performances

We begin with the evaluation of the in-control state. Figure 10 illustrates the empirical false alarm rates for all considered multivariate control charts in each of the 72 examined scenarios for all three applied control chart types. The results largely meet our expectations as the parametric Hotelling T^2 procedure tends to yield relatively low control limits. This particularly seems to be the case for more unbalanced setups (e.g. C_mD_mS, $B_dC_dD_d$, SAL) which tend to generate more skewed joint distributions as explained in Section 3.3.2. Overall, however, the desired fit is not reached for more balanced sets as well, since their false alarm rates lie above the desired α and above the ones of the other two procedures. See Figure 11 for an example of the quality of the fit where the deviation of the empirical distribution to the assumed F assumption is clearly visible, particularly at the tails of the distribution. Regarding the other control charts, the results support the statement that MEWMA is more robust against possible parametric violations. However, the non-parametric bootstrap approach clearly yields the most reliable in-control results for the rather long in-control phase and the small value of α.

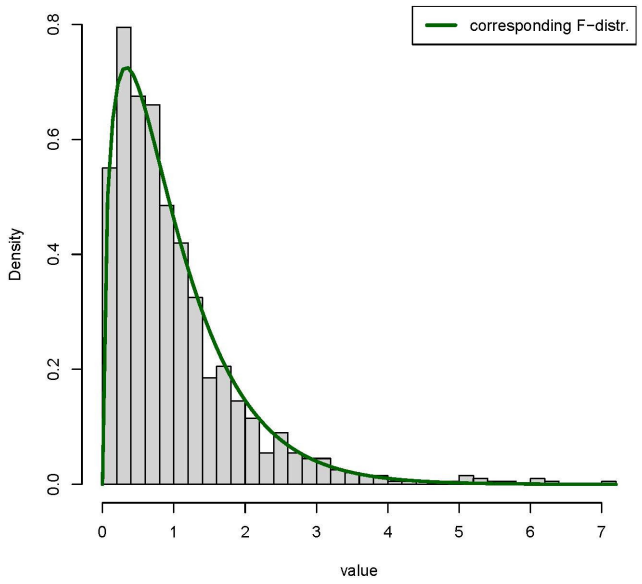


Figure 11: Example situation for a comparison of the simulated empirical distribution with the corresponding F-approximation of the parametric Hotelling T^2 chart.

3.4.3 Phase II Performances

We saw that the charts produce quite different ARL_0 values, although they were designed to hold a common fixed false alarm rate. Obviously, charts with a lower ARL_0 will produce lower values of ARL_1 on the same data set, since the control limit h_u is lower. Hence, the ARL_1 results should only be compared between the same applied control chart type. As the bootstrap chart achieved the most reliable Phase I performances, we therefore report the Phase II results only for the bootstrap chart and concentrate on the performance differences of the applied metric sets. See Table 14 for an overview of the results for the univariate procedure and the representatives of the multivariate monitoring strategies. The results for all 84 triple-combinations are given in Tables 17–23 in the Appendix.

Overall, the results meet our expectations. The univariate metrics might perform reasonably in special scenarios, but show clear weaknesses in others. The multivariate procedures, on the other hand, perform clearly more reliable over all situations and are more robust to the underlying change-type which underlines the improved flexiblity compared to the univariate approach. On a further note, the results are handily interpretable and are aligned with the statements of our derived categorization of Section 3.3.1. We can e.g. take a closer look at the performance of a representative of Category I, i.e the balanced set SB_dE (consisting of S, B_d, and E). Regarding the involved univariate performances, S is able to handle global changes, but has problems with LLCs and especially LNCs. For B_d, the behavior is vice versa.

Table 14: ARL_1 for the studied change types. Best ARL_1 for each scenario across all setups (univariate and multivariate) is printed in bold.

(a) Univariate Setup Performances

Type	Intensity	F	S	C_m	C_d	D_m	D_d	B_m	B_d	E
GLC	small	3.12	3.24	**3.03**	94.78	3.12	109.89	3.17	55.63	50.61
	moderate	1.22	**1.20**	**1.20**	68.86	1.23	86.83	1.22	16.89	19.82
	heavy	**1.00**	**1.00**	**1.00**	45.02	**1.00**	102.13	**1.00**	4.71	9.82
LLC	small	9.19	**7.70**	8.40	25.22	9.19	27.44	9.53	60.99	36.00
	moderate	14.57	9.97	12.14	3.14	14.57	3.45	14.57	6.04	4.43
	heavy	22.52	11.60	17.67	**1.41**	32.52	1.46	21.72	2.27	1.81
GNC	small	144.02	2.19	1.41	45.33	2.17	78.00	**1.34**	8.16	11.98
	moderate	142.55	1.01	**1.00**	9.82	1.01	73.06	**1.00**	2.02	3.03
	heavy	135.90	**1.00**	**1.00**	1.47	**1.00**	61.49	**1.00**	1.34	1.06
LNC	small	93.22	39.53	1.06	1.16	9.01	1.07	1.80	1.07	**1.00**
	moderate	90.57	95.01	4.45	1.02	38.66	**1.00**	4.46	1.07	**1.00**
	heavy	88.24	93.16	23.84	**1.01**	62.04	**1.01**	13.16	1.02	**1.01**

(b) Multivariate Setup Performances

Type	Intensity	SB_dE	SD_mC_d	B_dD_dS	C_mD_mS	$B_dC_dD_d$	FSB_d	SAL
GLC	small	20.56	15.63	19.00	8.90	25.31	23.52	41.34
	moderate	2.22	1.82	2.25	1.66	6.28	2.39	10.61
	heavy	**1.00**	**1.00**	**1.00**	**1.00**	1.04	**1.00**	1.06
LLC	small	27.72	18.47	25.47	68.36	22.07	29.05	23.56
	moderate	4.66	3.49	4.07	13.81	**2.92**	4.76	2.94
	heavy	1.86	1.60	1.76	17.27	1.52	2.04	1.51
GNC	small	7.49	12.38	6.87	6.06	8.88	23.94	7.09
	moderate	1.48	1.50	1.34	1.17	1.61	5.04	1.44
	heavy	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
LNC	small	1.02	1.03	1.03	1.00	1.02	1.50	1.11
	moderate	1.02	1.05	1.04	1.11	1.13	1.81	1.10
	heavy	**1.01**	**1.01**	**1.01**	2.00	**1.01**	**1.01**	**1.01**

Their joint monitoring, however, leads to promising results for all scenarios, since their effects are combined. The additional consideration of E, which is in theory sensitive to all types, but to a lesser extent, serves slightly supportive to all impacts and as an overall smoothing factor. The behavior for more unbalanced multivariate setups of Categories IIa, IIb, IIIa, IIIb is similar as they provide more flexibility overall compared to the involved univariate candidates. Particularly, the ARL_1 for the “non-specialized" cases (i.e. the cases, for which all involved metrics are not really suitable) improved as the small sensitivities of the single metrics have a larger impact if they are considered jointly, see e.g. the $B_dC_dD_d$ performance for GNCs or the C_mD_mS performance for LLCs. Despite the improved performance compared to the univariate metrics, the values are obviously higher than those of more balanced setups in these situations. Moreover, it is somewhat surprising that they also do not clearly outperform the more balanced sets in “specialized" cases, for which they are mainly constructed.

Lastly, the results for the performance of SAL underline our previous interpretations as this metric set achieves reasonable results for local changes, but struggles more in global change

Table 15: SBM Paramter Setups for the Community Changes

No.	In-Control	Changes
1	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.8, p_{ij} = 0.1,$ $n_1 = n_2 = 33, n_3 = 34$	$p_1 = 0.9, p_2 = 0.7,$ $p_3 = 0.9, p_{ij} = 0.2,$ $n_1 = n_2 = n_3 = 40$
2	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.8, p_{ij} = 0.1,$ $n_1 = n_2 = 33, n_3 = 34$	$p_1 = 0.4, p_2 = 0.9$
3	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.3, p_{ij} = 0.1,$ $n_1 = 30, n_2 = 20,$ $n_3 = 50$	$K = 4, p_4 = 0.3,$ $n_3 = 30, n_4 = 20$
4	$K = 3, p_1 = 0.7, p_2 = 0.6,$ $p_3 = 0.3, p_{ij} = 0.1,$ $n_1 = 30, n_2 = 20,$ $n_3 = 50$	$K = 4, p_3 = p_4 = 0.49$ $n_3 = 30, n_4 = 20$

cases compared to other multivariate sets. This behaviour underlines our classification into Category IIIb of Section 3.3.1. Furthermore, the additional consideration of the network density as a fourth metric compared to the set $\mathrm{B}_d\mathrm{C}_d\mathrm{D}_d$ seems to not contribute relevant added value as the performances are quite similar.

3.4.4 Extension to mixed-type changes

While this examination under rather rigid settings gave us crucial insights on the theoretical performance limits of the considered monitoring applications, we now examine if the studied behaviors still hold in practically more relevant application examples. For this purpose, we consider mixed-type changes (MTC) in the popular scenario of community changes in stochastic blockmodels (SBM) [Holland et al., 1983]. The explicit setups can be found in Table 15, where K represents the number of communities, p_i the intra-group link probability of group $i \in \{1, ..., K\}$, p_{ij} the inter-group link probability between groups i and j, and n_i the number of nodes in group i. Furthermore, we reduce the length of the in-control phase to 100 in order to examine the performance of the charts in more data-restrictive circumstances. Due to the shorter in-control length, we set the false alarm rate to $\alpha = 5\%$.

The in-control results shown in Figure 12 are different to those before. Whereas the non-parametric version clearly outperformed the parametric control charts in Section 3.4.2, the performances are more equal now. The non-parametric approach suffers from the shorter in-control length, because the estimation of the theoretical distribution becomes more unreliable by applying a smaller data sample to the bootstrap procedure. However, the chart still performs reasonably as it only lies approx. 0.5% above the desired α. Another advantage is the robustness against different metric sets. Overall, the parametric control charts perform better than before

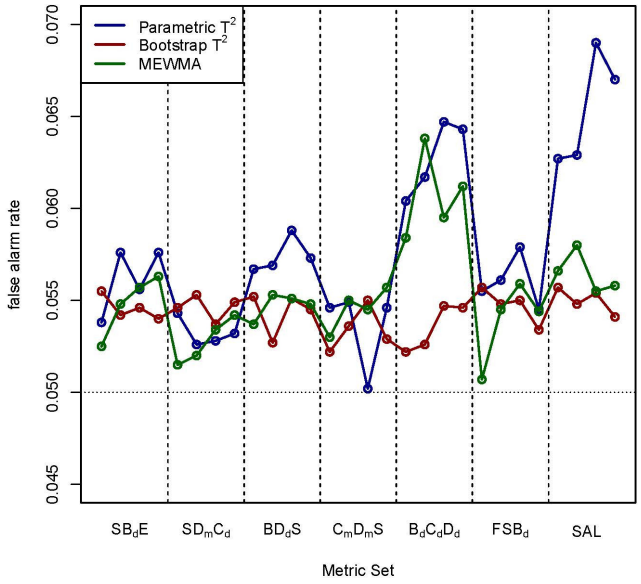


Figure 12: Average empirical false alarm rates for all simulated situations.

and reach similar performances for more balanced metric sets to the non-parametric candidate. An explanation is the higher value of α, for which the charts are designed, as explained in Section 3.3.2. However, the sensitivity to the applied metric set still holds as the performance gets quite unstable for unbalanced sets.

The ARL_1 results in Table 16 can be interpreted similarly as before. Siuation 1 describes a change with an increased popularity of the whole network with an increased inter- and intra-communication (link amount) and the arrival of new members which can be seen as a MTC of GNC and GLC. Apart from some univariate deviation centralities, all considered metric sets perform well. The second scenario addresses changes, where the importance between two groups is shifted, which results in an increased communication of one group and a decreased communication in the other. While most univariate metrics are not able to handle this change type as their ARL_1 does match their set ARL_0, the multivariate sets perform reasonably, particularly the more balanced ones. The last two situations address changes in the number of communities, e.g. a split of one group into two different ones. For the third situation, the link probability in both new groups stays the same as before which results in an overall decreased link amount due to the decreased link probability of those nodes, which were in one group before and are in different ones afterwards. Hence, it is a relatively easy GLC situation and all applied metrics achieve a satisfactory performance. We designed the last scenario such that the overall link probability stays the same after the split. This makes the detection more challenging which results in higher ARL_1 values. However, the multivariate sets again

Table 16: ARL_1 for the Community Changes. Best ARL_1 for each scenario across all setups is printed in bold.

(a) Univariate Setup Performances

Case	F	S	C_m	C_d	D_m	D_d	B_m	B_d	E
1	**1.00**	**1.00**	**1.00**	19.32	**1.00**	10.63	**1.00**	1.49	1.06
2	20.34	15.71	17.90	18.59	18.34	22.77	19.42	12.37	**1.62**
3	1.01	3.02	**1.00**	2.55	**1.00**	6.13	**1.00**	5.89	1.28
4	25.60	17.33	10.61	16.46	25.60	17.93	10.44	18.50	20.03

(b) Multivariate Setup Performances

Case	SB_dE	SD_mC_d	B_dD_dS	C_mD_mS	$B_dC_dD_d$	FSB_d	SAL
1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
2	1.87	5.00	9.87	5.12	14.58	4.77	13.75
3	1.39	**1.00**	2.21	**1.00**	1.43	**1.00**	1.03
4	8.94	17.18	14.14	**5.12**	10.74	15.25	10.92

underline their superior flexibility as they achieve better performances in this situation than their univariate counterparts.

3.5 Empirical Data Examples

To further underline the applicability and the handling of the approach, we analyze two real-world dynamic networks from economics and social sciences.

3.5.1 International Trade Data

This publicly available data set from the World Trade Organization (WTO) contains all reported international import-export relationships that are responsible for 90% of the overall worldwide trade volume. The data is collected quarterly and contains the time span of Q1 2010 - Q1 2022 which results in a dynamic network of $T = 49$ time points. For the selection of the in-control phase, we use the first 22 time points, i.e. the time span Q1 2010 - Q2 2015. We are not aware of major trading conflicts in this time period and therefore expect stable behaviour. As this in-control phase is rather short, we use the bootstrap Hotellings T^2 chart, because it promises to yield the most reliable results in these circumstances compared to the other charts as we saw in the simulation study. We do not expect a particular change type, since changes triggered by nodes and links both in a global and in a local fashion are conceivable. Therefore, we use the balanced metric set SB_dE in order to increase the probability for the detection of arbitrary change types. The result is shown in the upper part of Figure 13. The procedure detects a change in Q2 2020 that is probably triggered by the corona pandemic. Many countries were forced to reduce their trading activities and mainly restrict them to

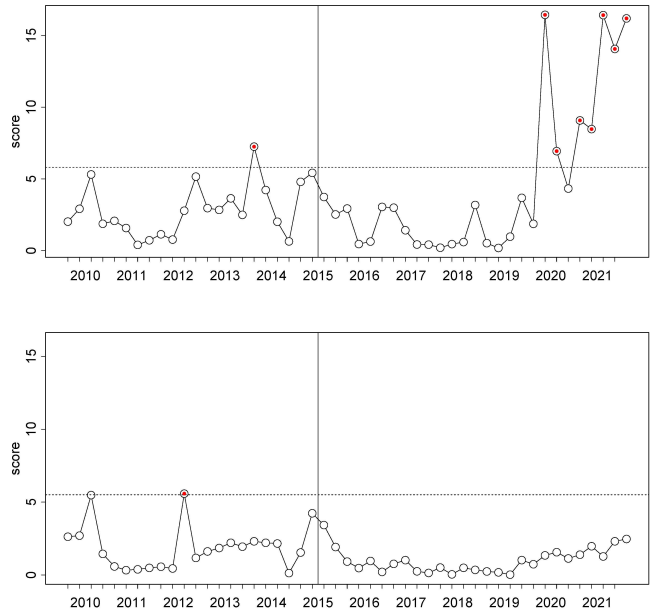


Figure 13: Control Charts using a balanced metric set (upper chart) and an unbalanced metric set (lower chart). In-control phase ends in Q2 2015. Red dots signalize alarms.

neighboring states and the most important partner nations. This behavior pushed the trade network to a more centralized layout with leading trade nations as the hubs. Consequently, we can classify the observed change as a local change. Interestingly, we can not detect global changes for this data as more unbalanced metric sets, that are focussed on global change types, do not trigger an alarm. This is visualized in the lower control chart of Figure 13, for which we used the metric set $C_m D_m S$. Hence, the pandemic apparently did not have a huge influence on the amount of relationships that are responsible for 90% of the overall worldwide trade volume, but it did change the network on a structural level as explained. This example again underlines the importance of a careful execution of the procedure and the flexibility advantage of a balanced metric set.

3.5.2 Enron Email Data

This data contains the email communication network between employees of the former US company Enron that has been made public by the US Department of Justice and was rigorously analyzed in many publications [Chapanond et al., 2005; Park et al., 2009; Priebe et al., 2005]. It comprises a time span from January 2000 until April 2002 on a monthly level. This results in $T = 28$ time points, from which we use the first 17 for the in-control phase and we aim to detect the effects of the accounting fraud scandal in the late 2001s and early 2002s. We again use the unbalanced metric set $SB_d E$ in a bootstrap Hotellings T^2 chart in order to be flexible regarding the existence of different types of change. The output of the resulting control chart is

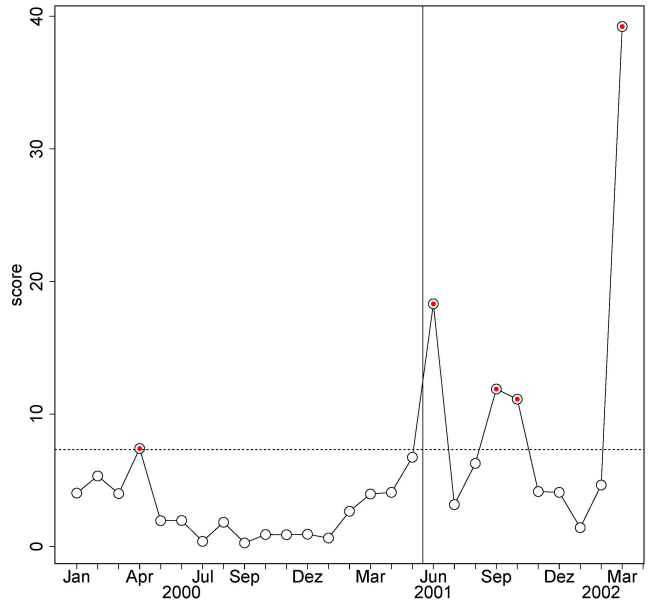


Figure 14: Control Chart with SB_dE for the Enron data. In-control phase ends in May 2001. Red dots signalize alarms.

depicted in Figure 14. The procedure triggers alarms that are in accordance with other social network analyses of this data set [Kendrick et al., 2018]. In June 2001 the CEO Jeffrey Skilling tried to fire the chairman and former CEO Kenneth Lay, in September/October 2001 Enron reported a huge loss and it was announced that a SEC (Security and Exchange Commission) inquiry has become a formal investigation, and in February/March 2002 more and more details of the fraud were made public.

3.6 Conclusion

The detection of temporal differences in a time series of graphs is a rather challenging task due to the complex nature of dynamic networks. We proposed an extension of a metric-based approach to a multivariate setup and its combination with suitable control charting procedures involving parametric as well as non-parametric setups. We explicitly explained the challenges of such a multivariate design and presented recommendations including a sound choice of a suitable set of metrics, its combination with a suitable control chart, and the final interpretation of the results. We particularly recommend to use a balanced metric set like SB_dE together with a non-parametric control chart like the bootstrap Hotelling T^2 chart in order to achieve reliable results in flexible change situations. We further validated our statements with the help of a simulation study and some real-world examples in which a thoroughly designed multivariate approach outperforms the univariate procedure by offering

a more flexible solution to the problem of change detection in social network analysis. As a main part of this paper was studying and evaluating different multivariate network metric sets and their capability of capturing relevant network information, the results can be beneficial for further statistical analysis procedures. In our view, this especially yields for statistical testing procedures (e.g. goodness-of-fit, two-sample tests) for network data where test statistics (i.e. metrics) need to be derived that characterize the network structure as comprehensively as possible.

Furthermore, our results can be beneficial also for researchers that aim to derive network monitoring procedures for a certain application scenario, e.g. with pre-known network structures (like in model-based approaches). In this context, our results can be used as a benchmark method as it is particularly designed for detecting changes of more general form.

Appendix

Table 17: ARL_1 multivariate performances.

Type	Intensity	FSC_m	FSC_d	FSB_m	FSB_d	FSD_m	FSD_d	FSE	FC_mC_d	FC_mB_m	FC_mB_d	FC_mD_m	FC_mD_d	FC_mE
GLC	small	5.30	16.76	12.53	23.53	4.39	14.06	8.73	5.36	8.66	5.74	4.46	5.13	4.96
	moderate	1.42	2.07	1.83	2.39	1.26	1.87	1.69	1.41	1.74	1.43	1.27	1.39	1.36
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLC	small	34.63	18.69	74.03	29.05	13.74	17.94	16.28	28.43	19.79	44.36	17.51	29.25	34.47
	moderate	12.51	3.41	44.51	4.76	20.30	3.79	5.28	4.42	15.99	5.32	16.79	4.76	6.56
	heavy	15.05	1.60	25.55	2.04	31.21	1.74	1.99	1.84	8.44	2.38	45.68	2.00	3.13
GNC	small	1.48	3.74	1.81	23.94	2.60	2.99	2.47	13.70	1.36	7.62	1.36	6.64	4.17
	moderate	1.00	1.00	1.00	1.04	1.00	1.00	1.00	1.12	1.00	1.13	1.00	1.06	1.04
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LNC	small	1.01	1.00	1.00	1.00	1.50	1.00	1.00	1.00	1.01	1.00	1.09	1.00	1.00
	moderate	1.44	1.01	1.08	1.01	1.81	1.00	1.01	1.00	2.48	1.01	3.53	1.00	1.01
	heavy	3.12	1.01	1.91	1.01	1.86	1.01	1.01	1.01	12.40	1.01	14.85	1.01	1.01

Table 18: ARL_1 multivariate performances.

Type	Intensity	FC_dB_m	FC_dB_d	FC_dD_m	FC_dD_d	FC_dE	FB_mB_d	FB_mD_m	FB_mD_d	FB_mE	FB_dD_m	FB_dD_d	FB_dE	FD_mD_d
GLC	small	16.04	30.75	4.38	35.88	17.99	15.71	5.33	14.72	14.45	4.39	28.35	25.61	4.23
	moderate	1.91	2.51	1.26	10.72	2.02	2.10	1.39	1.84	1.83	1.26	2.18	2.21	1.26
	heavy	1.00	1.00	1.00	1.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLC	small	45.83	26.73	13.54	23.00	20.67	105.27	18.46	71.36	107.86	15.00	28.36	29.95	13.66
	moderate	7.43	3.60	9.59	3.05	3.20	12.15	36.01	9.89	42.68	12.64	4.23	4.53	12.18
	heavy	2.11	1.60	2.72	1.52	1.54	2.66	52.74	2.39	4.55	3.35	1.77	1.93	3.30
GNC	small	2.82	6.72	3.96	11.79	4.37	3.14	1.38	2.49	2.17	3.83	5.47	7.65	2.84
	moderate	1.00	1.15	1.00	1.59	1.03	1.00	1.00	1.00	1.00	1.03	1.15	1.78	1.00
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LNC	small	1.00	1.02	1.00	1.00	1.04	1.00	1.05	1.00	1.00	1.00	1.02	1.02	1.00
	moderate	1.00	1.04	1.00	1.06	1.04	1.00	5.77	1.00	1.01	1.00	1.03	1.02	1.00
	heavy	1.01	1.01	1.01	1.01	1.01	1.01	33.69	1.01	1.01	1.01	1.01	1.01	1.01

Table 19: ARL_1 multivariate performances.

Type	Intensity	FD_mE	FD_dE	SC_mC_d	SC_mB_m	SC_mB_d	SC_mD_m	SC_mD_d	SC_mE	SC_dB_m	SC_dB_d	SC_dD_m	SC_dD_d	SC_dE
GLC	small	4.30	13.64	13.05	4.37	16.80	8.90	11.86	8.82	17.46	22.59	15.63	32.38	14.90
	moderate	1.26	1.85	1.74	1.33	1.83	1.66	1.69	1.54	1.87	2.31	1.82	8.48	2.11
	heavy	1.00	1.00	1.00	1.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.00
LLC	small	14.40	18.96	24.54	11.48	45.51	68.36	27.85	29.39	23.02	24.58	18.47	22.99	17.41
	moderate	19.61	3.32	3.50	8.22	4.59	13.81	3.96	6.19	3.49	3.52	3.49	3.00	3.16
	heavy	14.55	1.67	1.68	6.38	2.12	17.27	1.82	2.30	1.76	1.62	1.60	1.49	1.52
GNC	small	2.51	2.91	16.80	1.36	8.75	6.06	8.08	5.68	2.81	7.76	12.38	14.03	5.84
	moderate	1.00	1.00	1.48	1.00	1.37	1.17	1.24	1.14	1.00	1.52	1.50	2.00	1.18
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LNC	small	1.00	1.04	1.00	1.41	1.00	1.00	1.00	1.00	1.00	1.03	1.00	1.00	1.06
	moderate	1.01	1.02	1.00	12.48	1.00	1.11	1.00	1.02	1.00	1.05	1.03	1.06	1.06
	heavy	1.01	1.01	1.01	49.96	1.01	2.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01

Table 20: ARL_1 multivariate performances.

Type	Intensity	SB_mB_d	SB_mD_m	SB_mD_d	SB_mE	SB_dD_m	SB_dD_d	SB_dE	SD_mD_d	SD_mE	SD_dE	$C_mC_dB_m$	$C_mC_dB_d$	$C_mC_dD_m$
GLC	small	28.69	99.70	15.27	10.08	22.25	19.00	20.56	13.13	9.47	12.23	5.86	20.42	6.75
	moderate	2.29	12.37	1.84	1.64	2.15	2.25	2.22	1.79	1.60	1.85	1.59	2.08	1.66
	heavy	1.00	1.01	1.00	1.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLC	small	43.46	100.17	27.04	27.51	31.81	25.47	27.72	17.70	16.64	16.29	13.68	23.80	36.03
	moderate	4.76	135.18	3.98	5.93	4.69	4.07	4.66	3.69	5.20	3.27	3.66	3.52	3.61
	heavy	2.13	32.76	1.82	2.27	2.04	1.76	1.86	1.72	1.97	1.65	1.79	1.62	1.83
GNC	small	2.84	2.26	2.40	2.00	8.39	6.87	7.49	9.93	8.40	3.43	1.53	8.85	16.30
	moderate	1.00	1.00	1.00	1.00	1.57	1.34	1.48	1.42	1.34	1.04	1.00	1.62	1.47
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LNC	small	1.00	1.00	1.00	1.00	1.01	1.03	1.02	1.00	1.00	1.06	1.00	1.00	1.00
	moderate	1.00	1.08	1.00	1.01	1.02	1.04	1.02	1.01	1.01	1.02	1.00	1.01	1.00
	heavy	1.01	1.81	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01

Table 21: ARL_1 multivariate performances.

Type	Intensity	$C_mC_dD_d$	C_mC_dE	$C_mB_mB_d$	$C_mB_mD_m$	$C_mB_mD_d$	C_mB_mE	$C_mB_dD_m$	$C_mB_dD_d$	C_mB_dE	$C_mD_mD_d$	C_mD_mE
GLC	small	33.35	12.82	7.21	25.38	5.74	5.52	8.75	17.13	15.08	6.26	5.99
	moderate	8.33	1.88	1.63	2.76	1.47	1.44	1.74	1.90	1.81	1.56	1.52
	heavy	1.02	1.01	1.00	1.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LLC	small	22.94	16.79	18.21	23.52	14.35	12.59	81.27	26.80	26.55	39.48	63.41
	moderate	3.03	3.13	4.86	24.81	4.08	5.23	4.95	4.07	4.45	4.13	5.05
	heavy	1.52	1.54	2.10	8.61	1.87	2.49	2.53	1.77	1.94	2.01	3.17
GNC	small	12.60	5.76	1.60	1.36	1.52	1.48	8.90	7.79	8.61	8.29	5.84
	moderate	2.85	1.26	1.00	1.00	1.00	1.00	1.39	1.38	1.49	1.24	1.14
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LNC	small	1.16	1.00	1.00	1.02	1.01	1.03	1.02	1.00	1.00	1.06	1.00
	moderate	1.08	1.00	1.00	2.66	1.00	1.01	1.01	1.01	1.07	1.00	1.02
	heavy	1.01	1.01	1.01	12.31	1.01	1.01	1.01	1.01	1.01	1.01	1.01

Table 22: ARL_1 multivariate performances.

Type	Intensity	C_mD_dE	$C_dB_mB_d$	$C_dB_mD_m$	$C_dB_mD_d$	C_dB_mE	$C_dB_dD_m$	$C_dB_dD_d$	C_dB_dE	$C_dD_mD_d$	C_dD_mE	C_dD_dE
GLC	small	8.94	23.41	104.23	37.26	14.70	23.39	25.31	43.36	36.64	14.09	19.58
	moderate	1.72	2.20	13.19	9.82	2.00	2.09	6.28	9.33	9.71	2.00	4.58
	heavy	1.00	1.00	1.01	1.07	1.00	1.00	1.03	1.13	1.05	1.00	1.01
LLC	small	17.02	27.48	47.80	22.92	20.63	24.81	22.07	27.40	22.76	18.66	17.89
	moderate	3.27	3.59	10.05	3.03	3.19	3.59	2.92	3.29	3.03	3.19	2.99
	heavy	1.62	1.61	2.06	1.52	1.57	1.59	1.52	1.60	1.52	1.54	1.50
GNC	small	4.12	4.46	2.73	4.06	4.45	7.73	8.88	5.81	13.90	5.48	4.09
	moderate	1.07	1.05	1.00	1.08	1.04	1.56	1.61	1.16	2.00	1.18	1.12
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LNC	small	1.00	1.00	1.00	1.00	1.00	1.04	1.02	1.07	1.00	1.07	1.00
	moderate	1.00	1.00	1.00	1.05	1.03	1.05	1.13	1.07	1.06	1.06	1.08
	heavy	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01

Table 23: ARL_1 multivariate performances.

Type	Intensity	$B_mB_dD_m$	$B_mB_dD_d$	B_mB_dE	$B_mD_mD_d$	B_mD_mE	B_mD_dE	$B_dD_mD_d$	B_dD_mE	B_dD_dE	D_mD_dE
GLC	small	109.54	23.02	20.11	118.34	147.58	11.09	21.38	19.62	103.25	10.65
	moderate	17.90	2.03	1.94	13.49	13.14	1.77	2.03	1.94	23.57	1.77
	heavy	1.01	1.00	1.00	1.01	1.01	1.00	1.00	1.00	1.33	1.00
LLC	small	102.06	29.13	30.14	66.03	113.64	19.25	26.42	26.96	35.14	17.40
	moderate	17.35	4.21	4.55	17.86	106.32	3.29	4.16	4.50	4.22	3.27
	heavy	2.67	1.81	1.94	2.35	4.70	1.69	1.77	1.93	1.82	1.67
GNC	small	2.88	4.28	3.68	2.34	2.01	3.35	6.80	8.04	3.96	3.36
	moderate	1.00	1.02	1.02	1.00	1.00	1.01	1.33	1.44	1.03	1.44
	heavy	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
LNC	small	1.00	1.00	1.00	1.00	1.00	1.00	1.03	1.03	1.06	1.04
	moderate	1.00	1.00	1.00	1.00	1.01	1.01	1.04	1.02	1.06	1.02
	heavy	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01

4

Goodness-of-fit testing based on graph functionals for homogeneous Erdös-Rényi graphs

Based on: Brune, Barbara, Jonathan Flossdorf & Carsten Jentsch (2024): "Goodness-of-fit testing based on graph functionals for homogeneous Erdös-Rényi graphs", *Scandinavian Journal of Statistics*, 1–49, DOI:10.1111/sjos.12750.

Abstract

The Erdös-Rényi graph is a popular choice to model network data as it is parsimoniously parameterized, straightforward to interpret and easy to estimate. However, it has limited suitability in practice, since it often fails to capture crucial characteristics of real-world networks. To check its adequacy, we propose a novel class of goodness-of-fit tests for homogeneous Erdös-Rényi models against heterogeneous alternatives that permit non-constant edge probabilities. We allow for both asymptotically dense and sparse networks. The tests are based on graph functionals that cover a broad class of network statistics for which we derive limiting distributions in a unified manner. The resulting class of asymptotic tests includes several existing tests as special cases. Further, we propose a parametric bootstrap and prove its consistency, which avoids the often tedious variance estimation for asymptotic tests and enables performance improvements for small network sizes. Moreover, under certain fixed and local alternatives, we provide a power analysis for some popular choices of subgraph counts as goodness-of-fit test statistics. We evaluate the proposed class of tests and illustrate our theoretical findings by simulations.

4.1 Introduction

Due to the technical progress in recent decades, not only the amount of data is growing but also its complexity. Hence, the analysis of statistical network data has become a popular research objective with applications ranging from social sciences [Carrington et al., 2005; Sarkar and Moore, 2005] or biology [Bassett and Sporns, 2017; Prill et al., 2005] to logistics and transportation processes [Lee and Dong, 2009]. Various mathematical and computational methods have been developed to analyze, model, and understand the behavior of networks. As network data is rather complex by nature, it is a typical first step to employ suitable complexity reduction methods. A common approach is to calculate various sorts of summary statistics from the observed network such as e.g. matrix norms or centrality measures that may describe the network structure from different perspectives. However, the general difficulty is to be aware of the limited information that those statistics are able to capture, since they reduce the information of a whole network to a few scalar values [Flossdorf et al., 2023; Flossdorf and Jentsch, 2021; Ofori-Boateng et al., 2021]. Another approach is to fit a suitable network model to the observed data in order to analyze graphs in a simplified and controlled setup and to perform statistical inference. One of the simplest network models is the Erdös-Rényi model [Gilbert, 1959] in which the edges are considered as independent Bernoulli random variables with a common probability parameter p. Obviously, this model may have substantial shortcomings in modeling real world networks. Due to the assumption that all edges form independently with the *same* probability p, it usually fails to capture many of their features. For instance, in social networks, it is plausible that two people are more likely to know each other if they have a common friend. This yields clusters of vertices that have a higher edge probability between each other than to other vertices. That is, there may be several groups in a network, where two members of the same group have a higher probability to be connected than two members from different groups. A possible solution for this is to fit more general random graph models that allow for varying edge probabilities leading to the class of *heterogeneous* Erdös-Rényi models [Ouadah et al., 2020] including, e.g., the popularly used stochastic block model (SBM) [Holland et al., 1983] that can also be seen as a special case of graphons [Gao et al., 2015; Lovász, 2012] . Whereas these heterogeneous models might match the features of real world networks noticeably better than their homogeneous counterparts, their rigorous analysis is much more cumbersome from a graph theory point of view. From a practitioner's perspective, it is desirable to choose a parsimonious model with as few parameters as possible that still leads to a good fit to the data in order to simplify the estimation process. Hence, the homogeneous Erdös-Rényi model with only one probability parameter is still a popular benchmark model. Therefore, it is a crucial question if the usage of a parsimonious model is justified or if a more complex model achieves a significantly better fit. This leads to the field of goodness-of-fit procedures that aims to decide whether a specified model fits the underlying data adequately or if a more complex alternative would be a better choice.

Several approaches for goodness-of-fit testing for random graphs have been proposed in the literature. This involves tests for the number of communities in stochastic block models which are based on the largest singular value of a residual matrix that is obtained by the difference of the estimated block mean effect and the adjacency matrix [Lei, 2016]. Jin et al., 2018 developed a testing procedure to decide whether an underlying network has more than one community by counting motifs under the consideration of the problem of degree heterogeneity. Under similar conditions, these authors later established optimality results for a class of tests that are based on signed polygon statistics [Jin et al., 2021]. Testing procedures for the existence of communities are also studied in Yuan et al., 2022 using hypergraph motifs. Furthermore, the principal eigenvalue [Bickel and Sarkar, 2016] or the maximum entry [Hu et al., 2021] of a centered and scaled adjacency matrix is used to test whether a random graph is generated by a homogeneous Erdös-Rényi model or by a stochastic block model. For the same hypotheses, Gao and Lafferty, 2017 developed an asymptotic test based on specific induced subgraph counts. Subgraph counts as a test statistic are also used in Maugis et al., 2020 under the rather restrictive and often unrealistic assumption that multiple independent and identically distributed samples of graphs of the same size are available, and in Ospina-Forero et al., 2019 within a Monte-Carlo framework. A more general goodness-of-fit method is proposed in Dan and Bhattacharya, 2020, where the goal is to test whether the probability matrix of independent samples of a heterogeneous graph matches a predetermined reference matrix. To do so, optimal minimax sample complexities in various matrix norms such as e.g. the Frobenius norm are derived. In Ouadah et al., 2020, goodness-of-fit tests for Erdös-Rényi-type models have been derived based on the degree variance statistic that serves as an heterogeneity index of a graph [Snijders, 1981]. A statistical test for exponential random graph models [Robins et al., 2007] is proposed in Xu and Reinert, 2021 with test statistics that are derived from a kernel Stein discrepancy. Furthermore, a goodness-of-fit method for stochastic actor oriented models [Snijders, 1996] is presented in Lospinoso and Snijders, 2019 and variants of third-order motifs are used for the alternative of latent space models [Bubeck et al., 2016].

In this paper, we investigate a unified approach for goodness-of-fit testing for homogeneous Erdös-Rényi models. Precisely, we propose tests that aim to tell whether an observed network was either generated by some homogeneous Erdös-Rényi model or by some heterogeneous alternative model. Contrary to the related literature, we do not limit our analysis to just one particular test statistic, but use the concept of graph functionals [see Janson et al., 2011] for goodness-of-fit testing which allows for a unified treatment in theory and practice. Precisely, we propose a class of test statistics that covers a wide range of popular network metrics including e.g. the degree variance statistic, average centrality metrics, as well as raw and centered subgraph counts of arbitrary order and shape among others. Hence, our approach is very flexible and contains various already proposed tests [Gao and Lafferty, 2017; Ouadah et al., 2020] as special cases. Following this approach, we derive general asymptotic theory

for the whole proposed class of test statistics in a unified manner. In particular, for newly proposed test statistics that fall into our framework, this avoids the tedious derivation of the individual asymptotics in a case by case manner. Furthermore, while the implementation of asymptotic tests is often tedious as their variances are highly case-dependent, we also propose a simple parametric bootstrap approach and prove a general bootstrap consistency result in order to develop bootstrap versions of the goodness-of-fit tests. This circumvents the tedious variance estimation and enables finite sample improvements.

The paper is organized as follows. It starts off with some general random graph theory and introduces the relevant notation in Section 4.2. In particular, we recap the concept of graph functionals for Erdös-Rényi models and discuss their representation as weighted sums of (centered) subgraph counts, which enables the derivation of general asymptotic theory for graph functionals. Based on this theory, we construct a novel class of asymptotic goodness-of-fit tests in Section 4.3. We discuss some examples of graph functionals, including e.g. the degree variance as well as general notions of centered and raw subgraph counts, and investigate their suitability as goodness-of-fit statistics. Moreover, we provide a power analysis for the popular choices of subgraph counts of triangles and two-stars under certain fixed and local stochastic block model alternatives. In Section 4.4, to avoid the often tedious variance estimation for the construction of asymptotic tests, we propose a general parametric bootstrap procedure for graph functional-based goodness-of-fit testing and establish bootstrap consistency results that justify the resulting bootstrap critical values. Subsequently, we investigate the performance of asymptotic and bootstrap versions of several graph functional-based goodness-of-fit tests in an extensive simulation study in Section 4.5. The final Section 4.6 consists of some concluding remarks. Proofs and additional results are deferred to the Appendix.

4.2 Preliminaries: Random Graphs and Graph Functionals

In this chapter, we give an overview of the used network generating models and the concept of graph functionals. Particularly, we gather existing limiting distribution theory for this class of statistics for Erdös-Rényi graphs in order to enable the development of unified asymptotic and bootstrap goodness-of-fit tests.

4.2.1 Settings

Suppose we observe a random graph G that is defined by a VERTEX SET $V = V(G)$, and an EDGE SET $E = E(G)$. The edge set consists of pairs of vertices $\{v_1, v_2\}$, where $v_1, v_2 \in V(G)$. We denote by $n := n(G) = |V(G)|$ the number of vertices, and by $m := m(G) = |E(G)|$ the number of edges of the graph G, respectively. The vertex set is denoted by $V(G) = \{v_1, ..., v_n\}$,

and each edge in $E(G) = \{e_1, ..., e_m\}$ connects two of the vertices. We define the empty graph as a graph with no edges and the null graph $\emptyset$ as the graph with no vertices and, thus, no edges. A network can alternatively be represented by an adjacency matrix $A = (A_{ij})$ of dimension $(n \times n)$. In this paper, we concentrate on unweighted and undirected networks without self-loops, which are often called simple graphs. Their adjacency matrices are symmetric with binary entries such that $A_{ij} = 1$ indicates the presence of an edge between two vertices i and j, and $A_{ij} = 0$ if there is no edge between the respective vertices. As self-loops are not allowed, we have $A_{ii} = 0$ for all i. Throughout the paper, we focus on random graphs and use the well-established Erdös-Rényi graph as the benchmark model.

Definition 1 ***[Erdös-Rényi (ER) graph]*** *Let $G = (V, E)$ be a random graph on n vertices with adjacency matrix A and let $p \in [0, 1]$ be the connection probability. Then, we call G an Erdös-Rényi graph $\mathcal{G}(n, p)$, if the edges are realizations of stochastically independent and identically Bernoulli distributed random variables. That is, for $1 \leq i < j \leq n$, we have $A_{ij} \sim \text{Bin}(1, p)$ with $A_{ji} := A_{ij}$, and $A_{ii} := 0$ for all i. We denote the resulting ER model class by*

$$\mathcal{G}_{ER}(n) = \{\mathcal{G}(n, p), p \in [0, 1]\}.$$

As the ER model has quite restrictive assumptions, various generalizations have been studied. A quite flexible alternative is the heterogeneous Erdös-Rényi model [Ouadah et al., 2020].

Definition 2 ***[Heterogeneous Erdös-Rényi (HER) graph]*** *Let $G = (V, E)$ be a random graph on n vertices with adjacency matrix A and let $\boldsymbol{P} = (p_{ij})_{i,j=1,...,n}$ be the symmetric $(n \times n)$ matrix of connection probabilities with $p_{ij} \in [0, 1]$. Then, we call G a heterogeneous Erdös-Rényi graph $\mathcal{G}(n, \boldsymbol{P})$, if the edges are realizations of stochastically independent Bernoulli random variables. That is, for $1 \leq i < j \leq n$, we have $A_{ij} \sim \text{Bin}(1, p_{ij})$ with $A_{ji} := A_{ij}$, and $A_{ii} := 0$ for all i. We denote the resulting* HER *model class by*

$$\mathcal{G}_{\text{HER}}(n) = \{\mathcal{G}(n, \boldsymbol{P}), \boldsymbol{P} = (p_{ij}), p_{ii} = 0, p_{ij} = p_{ji}, p_{ij} \in [0, 1] \ \forall \ i, j\}.$$

In a nutshell, the HER model extends the classical ER model by offering the opportunity for individual link probabilities for each edge. This expansion is helpful for the modeling of more flexible scenarios. However, it also increases the complexity. E.g., parameter estimation in this model becomes infeasible without further restrictions. Hence, we are interested in testing whether a parsimonious homogeneous ER model is already sufficient to model the underlying data or whether an HER model achieves a significantly better fit. Consequently,

having observed a simple graph G of size n, we consider the testing problem

$$\mathrm{H}_0 : \; G \in \mathcal{G}_{\mathrm{ER}}(n) \qquad \text{vs.} \qquad \mathrm{H}_1 : \; G \in \mathcal{G}_{\mathrm{HER}}(n) \backslash \mathcal{G}_{\mathrm{ER}}(n). \tag{4.1}$$

Although $\mathcal{G}_{\mathrm{HER}}(n)$ can be decomposed into disjoint sets $\mathcal{G}_{\mathrm{ER}}(n)$ and $\mathcal{G}_{\mathrm{HER}}(n) \backslash \mathcal{G}_{\mathrm{ER}}(n)$ as above, it will not be possible to consistently detect *arbitrary* alternatives as $n \to \infty$, when testing H_0 against H_1. This is because a heterogeneous ER model has to deviate *sufficiently enough* from the homogeneous model to be able to detect it, since we only rely on a single network observation in the goodness-of-fit context. For example, consider an HER model that has the same edge probability p for all except a finite and (for increasing n) fixed number of edges, that have edge probability q with $q \neq p$. Then, as only finitely many edge probabilities deviate from p, such heterogeneous alternatives will be not detectable in the limit as n increases.

4.2.2 Graph Functionals

We use the concept of graph functionals for the development of a general and rich class of goodness-of-fit tests. In this subsection, we gather existing theory for graph functionals that is required in order to derive general asymptotic theory.

General Concept

Graph functionals cover a wide range of network statistics that can capture various forms of structural patterns within a given network. Hence, the class of graph functionals appears to be favorable for goodness-of-fit testing. Following Janson et al., 2011, we make use of the following definition of graph functionals.

Definition 3 ***[Graph isomorphism, automorphism and graph functional]*** *Let G and H be two simple graphs. An isomorphism of graph G onto H is a bijection $\varphi : V(G) \to V(H)$ such that any two vertices v_1, $v_2 \in V(G)$ are adjacent in G if and only if the vertices $\varphi(v_1)$, $\varphi(v_2) \in V(H)$ are adjacent in H. If there exists an isomorphism between G and H, we call G and H isomorphic. A real-valued random variable $X_n = X_n(G)$ that only depends on the isomorphism type of a graph G of size n, i.e. with $X_n(G) = X_n(H)$ for all G and H that are isomorphic, is called a graph functional. Further, we denote by* $\mathrm{aut}(H)$ *the number of automorphisms of H, i.e., the number of isomorphisms for $G = H$.*

In other words, a graph functional is a function computed from a graph G that does not hinge on the vertex labels. We give some examples of special cases next, which will serve as running examples throughout this paper for illustration purposes.

Example 1. [**Frobenius norm**] A rather simple example for a graph functional is the squared Frobenius norm $\|A\|_F^2$ of the adjacency matrix $A = (A_{ij})$ of a graph $G = (V, E)$. It is defined by

$$\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2.$$

By summing up all (squared) entries of A, it is easy to see that relabeling the vertices of G does not change $\|A\|_F^2$ such that $\|A\|_F^2$ is a graph functional. As we concentrate on undirected and unweighted graphs, which have a symmetric adjacency matrix A with binary entries A_{ij}, we have

$$\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}^2 = 2 \sum_{1 \le i < j \le n} A_{ij} = 2m = 2\binom{n}{2}\widehat{p}, \tag{4.2}$$

where $m = |E(G)|$ is the number of edges of G and $\widehat{p} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} A_{ij}$ is the natural estimator of p based on the adjacency matrix $A = (A_{ij})$. Since the Frobenius norm counts the number of edges and is thus the same for every permutation of a given graph, it belongs to the class of graph functionals.

Example 2. [**Degree Variance**] The *degree variance* V_n of a (simple) graph $G = (V, E)$ is defined as

$$V_n := \frac{1}{n} \sum_{i=1}^{n} (D_i - \overline{D}_n)^2, \tag{4.3}$$

where D_i is the degree of vertex v_i and $\overline{D}_n = \frac{1}{n} \sum_{i=1}^{n} D_i$ denotes the average degree of the underlying network G. Note that the individual degrees D_i do actually hinge on the vertex labels, but the sum of (squared) degrees as well as the average degree do not change by relabeling the graph. Hence, V_n is invariant under isomorphism change and, consequently, a graph functional. The degree variance is quite popular as it is an intuitive metric that can handily be computed directly from the adjacency matrix. As pointed out by Snijders, 1981, it serves as a certain heterogeneity index of a graph and has been used for goodness-of-fit testing in Ouadah et al., 2020. Further investigated in Flossdorf and Jentsch, 2021, V_n performs reasonably well in capturing local structures of the graph (e.g. centrality traits), but also that V_n has weaknesses, when it comes to capturing global characteristics like the overall amount of links.

Example 3. [**Eigenvector Centrality**] Eigenvector centrality can be interpreted as a natural extension of the degree centrality. It is measured relatively to the amount of relations a node has to other nodes, and relationships to high-scoring nodes are more valuable than those to low-scoring ones. Thus, not only the number of links is taken into account but also the

influences of the nodes to which the links exist. By denoting the score of node i with x_i, the eigenvector centrality is formally defined by

$$x_i = \frac{1}{\lambda}\sum_{j=1}^{n} a_{ij}x_j, \quad i = 1, \ldots, n,$$

where λ is a constant. If we define the vector of scores as $x = (x_1, \ldots, x_n)$, it is possible to express the equation in matrix form $\lambda x = Ax$. Hence, λ is an eigenvalue of A and x is the corresponding eigenvector which includes the centrality scores. In order to get a metric for the whole network, the *average eigenvector centrality* is defined as

$$\mathrm{E}_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Note that the eigenvector values x_i are typically scaled to be able to compare the different scores properly with each other. We use the internal scaling of the eigenvector, that is, we replace x_i by $x_i/\|x\|$, as it has unit length in the Euclidean norm. By the same arguments as used in the above examples, the metric E_n is a graph functional.

While these examples focus on three concrete special cases of graph functionals, we consider also the more general concept of subgraph counts. According to Janson et al., 2011, p. 165, we define centered and raw versions of subgraph counts as follows.

Definition 4 ***[Subgraph counts]*** *Let G be a (simple) graph on n vertices, and H be another graph with $n(H) \leq n$. Consider the $(n)_{n(H)}$ different injective mappings φ from vertices of H into $\{1, ..., n\}$, where, for $x, y \in \mathbb{N}_0$ with $x \leq y$, let $(x)_y = x(x-1)(x-2)\cdots(x-y+1)$ denote the descending factorials. Each mapping φ maps H onto a copy $\varphi(H)$ of H in a complete graph with n vertices. Then, the centered subgraph count of H in G is defined by*

$$S_n(H) = \sum_{\varphi} \prod_{e \in E(\varphi(H))} \Big(\mathbb{1}\{e \in E(G)\} - \mathbb{P}(e \in E(G)) \Big),$$

while the raw subgraph count of H in G is defined by

$$T_n(H) := \frac{1}{\mathrm{aut}(H)} \sum_{\varphi} \prod_{e \in E(\varphi(H))} \mathbb{1}\{e \in E(G)\}.$$

It is worth noting that a more precise notation would be to use $S_G(H)$ and $T_G(H)$ as the subgraph counts depend on the whole graph G and not only on its size $n = n(G) = |V(G)|$. However, as we apply the subgraph counts always to a graph G, we suppress this dependence and write $S_n(H)$ and $T_n(H)$, respectively.

While, in a relabeled graph, the subgraph of interest may exist between different node combinations compared to the observed graph, the sum of such subgraphs remains the same over the whole network. Hence, centered subgraph counts $S_n(H)$ as well as raw subgraph counts $T_n(H)$ belong to the class of graph functionals and form important subclasses. Although defined similarly, following the expositions in Janson et al., 2011, they mainly differ in three important aspects:

(i) For $H \neq \emptyset$, $S_n(H)$ is a random variable with expectation 0, i.e., $E(S_n(H)) = 0$, for models with independently formed edges, as $\mathbb{P}(e \in E(G)) = \mathbb{E}(\mathbb{1}\{e \in E(G)\})$, which is not the case for $T_n(H)$, where $E(T_n(H)) > 0$ typically holds.

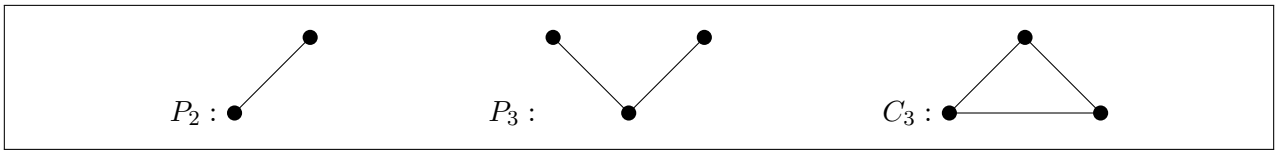
(ii) While $T_n(H)$ can be computed exclusively from the graph G (i.e., from its adjacency matrix A), this is not the case for $S_n(H)$, which requires knowledge of $\mathbb{P}(e \in E(G))$, i.e. p for Erdös-Rényi graphs.

(iii) While $T_n(H)$ counts a substructure H only once for each vertex combination (formally by dividing by $\mathrm{aut}(H)$), $S_n(H)$ counts all its automorphisms.

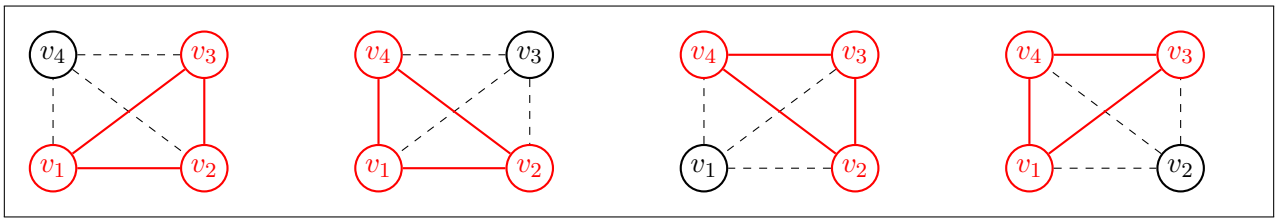
Subgraph counts are popular metrics for characterizing networks and are used for various inferential approaches [Bhattacharyya and Bickel, 2015; Gao and Lafferty, 2017; Maugis et al., 2020]. However, these methods typically use subgraph counts of a certain simple shape and small order, which makes them quite restrictive in their practical application. Even for simple subgraph structures, the computational complexity quickly becomes large compared to, for example, the degree variance statistic. However, as subgraph counts are always graph functionals, this property allows us to derive asymptotics in a unified way for flexible shapes and orders, as we will show throughout the paper. Popularly used subgraphs are e.g. triangles C_3 (△), two-stars P_3 (⩛) and single edges P_2 (／) which are depicted in Figure 15a.

Remark 1. [Differing notions of subgraph counts] Note that there are different notions of subgraph counts in the literature. Differences arise from considering the order of the vertices, i.e. counting a substructure once for each vertex combination, or counting all automorphisms. Another aspect is whether to consider only induced copies of subgraphs, i.e. also preserving non-edges, or all non-induced copies, not necessarily preserving non-edges. Our definitions above are in accordance with those of Janson et al., 2011. For $S_n(H)$, all automorphisms for each vertex combination are considered, as can be seen from the index of the sum. For $T_n(H)$, the prefactor causes us to count a substructure only once for each vertex combination, which is a somewhat more intuitive concept. Both $T_n(H)$ and $S_n(H)$ are defined as non-induced versions. However, under the condition $p = o(1)$ the induced and non-induced versions are asymptotically the same [Janson et al., 2011, p.78].

Figure 15b illustrates the mechanism of subgraph counts with an example: In a complete graph of order 4, we can find 4 subgraphs of shape C_3, hence $T_4(C_3) = 4$. For the calculation



(a) Examples for subgraphs: edge P_2, two-star P_3 and circle C_3 (of size 3).



(b) All raw subgraphs (circles) of shape C_3 in a complete graph with $n = 4$ vertices leading to $T_4(C_3) = 4$.

Figure 15: Illustration of the mechanism of subgraphs and subgraph counts.

of the centered versions, we would have to consider each of these $\text{aut}(C_3) = 6$ times, i.e. the sum in the definition of S_n would range over $(4)_3 = 4 \cdot 3 \cdot 2 = 24$ mappings φ leading then to $S_4(C_3) = 24$.

Unified Representation

The key foundation of our unified procedure for constructing goodness-of-fit tests lies in a fundamental relationship between graph functionals and *centered* subgraph counts: Any graph functional $X_n = X_n(G)$ can be written as a linear combination of centered subgraph counts $S_n(H)$. This expansion enables the unified derivation of the asymptotic distributions of graph functionals X_n and is referred to as the method of higher projections. In particular, the representation of X_n in terms of $S_n(H)$'s allows to exploit beneficial features of centered subgraph counts that are summarized in the following proposition.

Proposition 1 (Janson et al., 2011, Lemma 6.41) *Let H and J, $H, J \neq \emptyset$, be graphs without isolated vertices. Then, we have*

(a) $\mathbb{E}(S_n(H)) = 0$.

(b) $\mathbb{V}ar(S_n(H)) = aut(H)(n)_{n(H)}(p(1-p))^{m(H)}$, *where* $(x)_y = x(x-1)(x-2)\cdots(x-y+1)$ *denotes the descending factorials.*

(c) If H and J are non-isomorphic, then $S_n(H)$ and $S_n(J)$ are orthogonal, that is, $\mathbb{C}ov(S_n(H), S_n(J)) = 0$.

That is, for any non-null graph H, we have that $S_n(H)$ is a centered random variable, i.e. $\mathbb{E}(S_n(H)) = 0$ and its variance $\mathbb{V}\text{ar}(S_n(H))$ can be explicitly stated and has a rather simple form. Moreover, whenever two (non-null) graphs H and J are not isomorphic, then $S_n(H)$ and $S_n(J)$ are uncorrelated, i.e. $\mathbb{C}\text{ov}(S_n(H), S_n(J)) = 0$.

In addition to the nice properties summarized in Proposition 1, the following result states that centered subgraph counts $S_n(H)$ can be used to (linearly) decompose *any* graph functional.

Proposition 2 (Janson et al, 2011, Lemma 6.42) *Every graph functional X_n of a $\mathcal{G}(n,p)$ graph has a unique expansion*

$$X_n = \sum_{H \in \mathcal{H}} a_n(H) S_n(H) \tag{4.4}$$

for some are real-valued coefficients $a_n(H) = a_n(H,p)$ depending on n, p and H, where $\mathcal{H}$ contains all unlabelled graphs without isolated vertices and of at most n vertices. Furthermore, the terms in the sum are orthogonal, hence

$$\mathbb{V}ar(X_n) = \sum_{H \in \mathcal{H}} a_n^2(H) \mathbb{V}ar(S_n(H)). \tag{4.5}$$

Note that the expectation of X_n is represented solely by the coefficient $a_n(\emptyset)$, i.e. $E(X_n) = a_n(\emptyset)$. This is because the count of null graphs is deterministic and given by $S_n(\emptyset) := 1$ and all other terms $S_n(H)$, $H \neq \emptyset$, are centered by construction. While Proposition 2 guarantees the expansion (4.4) for arbitrary graph functionals X_n, it may be the case that not all centered subgraph counts $S_n(H)$ in (4.5) actually contribute (asymptotically) to $\mathbb{V}\text{ar}(X_n)$.

Definition 5 ***[Dominance of families of subgraphs]*** *For a family of non-null unlabelled graphs without isolated vertices $\mathcal{H}_0$, we call a graph functional X_n dominated by $\mathcal{H}_0$ if, for a given sequence of connection probabilities $p = p(n) \xrightarrow[n\to\infty]{} p_0 \in [0,1]$, it holds*

$$\frac{\mathbb{V}ar(X_n)}{\sum_{H \in \mathcal{H}_0} a_n^2(H) \mathbb{V}ar(S_n(H))} = \frac{\sum_{H \in \mathcal{H}} a_n^2(H) \mathbb{V}ar(S_n(H))}{\sum_{H \in \mathcal{H}_0} a_n^2(H) \mathbb{V}ar(S_n(H))} \xrightarrow[n\to\infty]{} 1.$$

Being dominated by a family of (sub)graphs $\mathcal{H}_0$ means that asymptotically all the variance of the graph functional X_n is explained by those centered subgraphs in the family $\mathcal{H}_0$, which may be smaller than $\mathcal{H}$ as defined in Proposition 2.

Asymptotics

The representation of graph functionals as linear combinations of centered subgraph counts enables a unified derivation of asymptotic theory for the whole class of graph functionals. While Nowicki and Wierman, 1988 and Nowicki, 1989 showed that raw subgraph counts $T_n(H)$ follow asymptotic normal distributions if the underlying model is contained in $\mathcal{G}_{\text{ER}}(n)$, we are interested in finding the asymptotic distributions for *centered* subgraph counts $S_n(H)$ in order to get also asymptotic results for graph functionals via Proposition 2. In particular, by combining the statements of Propositions 1 and 2 with the asymptotic results established for $S_n(H)$'s in Theorem 6.43 in Janson et al., 2011, we get the following general asymptotic normality result for graph functionals.

Theorem 1 (Janson et al., 2011, Theorem 6.49) *Let X_n be a graph functional of $\mathcal{G}(n,p)$ with $p = p(n) \xrightarrow[n\to\infty]{} p_0 \in [0,1]$. Suppose X_n is dominated by a family of connected graphs $\mathcal{H}_0$ such that, for all $H \in \mathcal{H}_0$ and $np^{r(H)} \xrightarrow[n\to\infty]{} \infty$ with $r(H) = \max_{J\subseteq H} d(J)$, where $d(J) = m(J)/n(J)$ denotes the density of J, the coefficients*

$$b(H) = \sup_n \frac{n^{n(H)/2} p^{m(H)/2} a_n(H)}{\sqrt{\mathbb{V}ar(X_n)}}$$

are finite and satisfy

$$\sum_{H\in\mathcal{H}_0} b(H)^2 aut(H) < \infty.$$

Then, as $n \to \infty$, it holds

$$R(X_n) := \frac{X_n - \mathbb{E}(X_n)}{\sqrt{\mathbb{V}ar(X_n)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{4.6}$$

Note that the expression $J \subseteq H$ in the above definition of $r(H)$ includes all graphs for which $E(J) \subseteq E(H)$ and $V(J) \subseteq V(H)$. As a direct consequence of Theorem 1, there exists a non-negative sequence $(c_n)_{n\in\mathbb{N}}$ with $c_n = c(n, p(n))$ such that $c_n^2 \mathbb{V}\text{ar}(X_n) \to V^2 > 0$ as $n \to \infty$, leading to $c_n(X_n - \mathbb{E}(X_n)) \xrightarrow{d} \mathcal{N}(0, V^2)$.

The general representation of graph functionals as linear combinations of centered subgraph counts in Proposition 2 and the limiting distributions derived in Theorem 1 form together a powerful framework that enables a unified treatment and a flexible construction of goodness-of-fit tests for (homogeneous) Erdös-Rényi random graphs.

4.3 Goodness-of-fit Testing for Erdös-Rényi Models

In this section, we formulate our unified approach for goodness-of-fit testing for ER graphs based on graph functionals. That is, having observed a simple graph G of size n and recalling (4.1), we restate the testing problem

$$\mathrm{H}_0 : \; G \in \mathcal{G}_{\mathrm{ER}}(n) \qquad \text{vs.} \qquad \mathrm{H}_1 : \; G \in \mathcal{G}_{\mathrm{HER}}(n) \backslash \mathcal{G}_{\mathrm{ER}}(n). \tag{4.7}$$

Consequently, under the assumptions of Theorem 1 and in view of the asymptotic normality result in (4.6), it appears natural to pick a certain graph functional $X_n = X_n(G)$ and to use $R(X_n)$ defined in (4.6) as test statistic for testing H_0 against H_1 in (4.7), where $\mathbb{E}(X_n)$ and $\mathbb{V}\mathrm{ar}(X_n)$ have to be replaced by $\mathbb{E}_{H_0}(X_n)$ and $\mathbb{V}\mathrm{ar}_{H_0}(X_n)$, which denote expectation and variance of X_n under H_0, respectively. However, this is usually not feasible, because neither $\mathbb{E}_{H_0}(X_n)$ nor $\mathbb{V}\mathrm{ar}_{H_0}(X_n)$ are typically known. Hence, instead, we will consider feasible goodness-of-fit test statistics of the form

$$\widehat{R}(X_n) := \frac{X_n - \widehat{\mathbb{E}_{H_0}}(X_n)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n)}}, \tag{4.8}$$

where $\widehat{\mathbb{E}_{H_0}}(X_n)$ and $\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n)$ are suitable (consistent) estimators of $\mathbb{E}_{H_0}(X_n)$ and $\mathbb{V}\mathrm{ar}_{H_0}(X_n)$, respectively. Such estimators are naturally obtained by replacing the unknown parameter p in the explicit expressions of $\mathbb{E}_{H_0}(X_n)$ and $\mathbb{V}\mathrm{ar}_{H_0}(X_n)$ by its consistent estimator $\widehat{p}$ defined in (4.2). However, note that $\mathbb{E}_{H_0}(X_n)$ and $\mathbb{V}\mathrm{ar}_{H_0}(X_n)$ are generally difficult to derive directly. Making use of Proposition 2 (see also the discussion below Proposition 2), under H_0, we have

$$\begin{aligned}
\mathbb{E}_{H_0}(X_n) &= \sum_{H \in \mathcal{H}} a_n(H) \mathbb{E}_{H_0}(S_n(H)) = a_n(\emptyset) \mathbb{E}_{H_0}(S_n(\emptyset)) = a_n(\emptyset), \\
\mathbb{V}\mathrm{ar}_{H_0}(X_n) &= \sum_{H \in \mathcal{H}} a_n^2(H) \mathbb{V}\mathrm{ar}(S_n(H)).
\end{aligned}$$

As the coefficients $a_n(H)$ depend on n, p and H, i.e. $a_n(H) = a_n(H, p)$, estimators $\widehat{\mathbb{E}_{H_0}}(X_n)$ and $\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n)$ are naturally obtained by defining

$$\begin{aligned}
\widehat{\mathbb{E}_{H_0}}(X_n) &= \widehat{a}_n(\emptyset), \\
\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n) &= \sum_{H \in \mathcal{H}} \widehat{a}_n^2(H) \widehat{\mathbb{V}\mathrm{ar}_{H_0}}(S_n(H)),
\end{aligned}$$

where $\widehat{a}_n(H) = a_n(H, \widehat{p})$ and $\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(S_n(H)) = \mathrm{aut}(H)(n)_{n(H)}(\widehat{p}(1-\widehat{p}))^{m(H)}$. With these prerequisites in place, by making use of Theorem 1, we are able to make the following statement about the limiting distribution of $\widehat{R}(X_n)$ defined in (4.8) under H_0.

Proposition 3 *Let $X_n = X_n(G)$ be a graph functional computed from an Erdös-Rényi graph G and suppose the assumptions of Theorem 1 are fulfilled. Additionally, let*

$$\text{(A)} \quad \frac{\mathbb{V}ar_{H_0}(X_n)}{\widehat{\mathbb{V}ar_{H_0}}(X_n)} \xrightarrow{P} 1 \quad \textit{and} \quad \text{(B)} \quad \frac{\mathbb{E}_{H_0}(X_n) - \widehat{\mathbb{E}_{H_0}}(X_n)}{\sqrt{\mathbb{V}ar_{H_0}(X_n)}} = \frac{a_n(\emptyset) - \widehat{a}_n(\emptyset)}{\sqrt{\mathbb{V}ar_{H_0}(X_n)}} \xrightarrow{P} 0 \tag{4.9}$$

hold. Then, as $n \to \infty$, we have

$$\widehat{R}(X_n) = \frac{X_n - \widehat{\mathbb{E}_{H_0}}(X_n)}{\sqrt{\widehat{\mathbb{V}ar_{H_0}}(X_n)}} \xrightarrow{d} \mathcal{N}(0,1).$$

An asymptotic level-α test for testing the null hypothesis H_0 of an underlying homogeneous Erdös-Rényi graph against heterogeneous alternatives H_1 in (4.1) *is given by the decision rule*

$$\phi_\alpha(X_n) = \begin{cases} 1, & \textit{if } \left|\widehat{R}(X_n)\right| > z_{1-\alpha/2}, \\ 0, & \textit{else,} \end{cases}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

The additional conditions (A) and (B) in (4.9) imposed in Proposition 3 are important to guarantee the limiting standard normal distribution of the *feasible* test statistic $\widehat{R}(X_n)$ in (4.8). While condition (A) is likely to hold in relevant cases of graph functionals used in practice, condition (B) is actually more delicate. As we will see in the following illustrating examples, it is case dependent, whether this condition holds for a certain graph functional or not. Next, we pick up Example 1 from above and show that the Frobenius norm $\|A\|_F^2$ is a graph functional, that does *not* satisfy property (4.9) as condition (B) fails to hold. Additionally, we discuss why $\|A\|_F^2$ is not suitable at all for testing H_0 against H_1.

Example 1 continued. [Frobenius norm] On the one hand, according to (4.2), we have $\|A\|_F^2 = \binom{n}{2}\widehat{p}$. On the other hand, we have $\widehat{p} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} A_{ij} = \binom{n}{2}^{-1} T_n(P_2)$ with

$$\begin{aligned} T_n(P_2) &= \frac{1}{\mathrm{aut}(P_2)} \sum_{\varphi} \prod_{e\in E(\varphi(P_2))} \mathbb{1}\{e \in E(G)\} = \frac{1}{\mathrm{aut}(P_2)} \sum_{\varphi} \prod_{e\in E(\varphi(P_2))} \Big(\mathbb{1}\{e \in E(G)\} - p + p\Big). \\ &= \frac{1}{\mathrm{aut}(P_2)} \left(\sum_{\varphi} \prod_{e\in E(\varphi(P_2))} \Big(\mathbb{1}\{e \in E(G)\} - p\Big) + \sum_{\varphi} \prod_{e\in E(\varphi(P_2))} p \right) \\ &= \frac{1}{\mathrm{aut}(P_2)} \left(S_n(P_2) + n(n-1)p\right) = \frac{1}{2} S_n(P_2) + \binom{n}{2} p, \end{aligned}$$

where we used that the product $\prod_{e\in E(\varphi(P_2))}$ consists only of one factor, as P_2 has only one edge, and $p = \mathbb{P}(e \in E(G))$. Altogether, when considering $\|A\|_F^2$ as a graph functional X_n, we have

the expansion

$$\|A\|_F^2 = \binom{n}{2}\widehat{p} = T_n(P_2) = \frac{1}{2}S_n(P_2) + \binom{n}{2}p = \frac{1}{2}S_n(P_2) + \binom{n}{2}pS_n(\emptyset). \tag{4.10}$$

That is, the (squared) Frobenius norm can likewise be expressed in terms of the estimated edge probability $\widehat{p}$, by using the raw subgraph count $T_n(P_2)$ or the centered subgraph count $S_n(P_2)$. According to Proposition 2, (4.10) gives exactly the representation of $X_n = \|A\|_F^2$ in terms of centered subgraph counts in (4.4) with $a_n(P_2) = \frac{1}{2}$ and $a_n(\emptyset) = \binom{n}{2}p$. Hence, according to Theorem 1, we have

$$\frac{\|A\|_F^2 - \mathbb{E}(\|A\|_F^2)}{\sqrt{\mathbb{V}\mathrm{ar}(\|A\|_F^2)}} = \frac{\frac{1}{2}S_n(P_2)}{\sqrt{\mathbb{V}\mathrm{ar}(\frac{1}{2}S_n(P_2) + \binom{n}{2}p^3)}} = \frac{S_n(P_2)}{\sqrt{\mathrm{aut}(P_2)n(n-1)p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1).$$

However, as we have

$$\frac{a_n(\emptyset) - \widehat{a}_n(\emptyset)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(\|A\|_F^2)}} = \frac{\binom{n}{2}p - \binom{n}{2}\widehat{p}}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(\binom{n}{2}\widehat{p})}} = \frac{\frac{1}{2}S_n(P_2)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(\frac{1}{2}S_n(P_2))}} \xrightarrow{d} \mathcal{N}(0,1)$$

again, condition (B) in (4.9) does *not* hold. Moreover, as $\|A\|_F^2 = \widehat{E_{H_0}}(\|A\|_F^2)$ holds, we even have

$$\widehat{R}(\|A\|_F^2) = \frac{\|A\|_F^2 - \widehat{\mathbb{E}_{H_0}}(\|A\|_F^2)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(\|A\|_F^2)}} = 0.$$

Consequently, for $X_n = \|A\|_F^2$ and under the null of an ER graph, in contrast to the infeasible test statistic $R(\|A\|_F^2)$, the feasible test goodness-of-fit test statistic $\widehat{R}(\|A\|_F^2)$ does not converge to a standard normal distribution. Actually, the feasible test statistic does also not converge to some other non-degenerate distribution, it becomes exactly zero such that this test statistic is not suitable at all for testing H_0 against H_1. But this makes a lot of sense, because just counting the number of edges in a graph G (this is what the (squared) Frobenius norm does), does not allow at all to distinguish between homogeneous and heterogeneous Erdös-Rényi graphs.

This example demonstrates that condition (B) can actually fail for (too) simple graph functionals. In the next subsection, we illustrate the whole machinery of graph functional-based goodness-of-testing and the derivation of asymptotic theory in detail for the example of the degree variance V_n defined in (4.3).

4.3.1 A deep example: Degree variance goodness-of-fit testing

Firstly, we would like to note that a goodness-of-fit test for V_n has already been proposed by Ouadah et al., 2020. Eventually, we will see that - although using a different concept of proof - we come to the same result for our asymptotic version which underlines the unified characteristic of our approach that contains this test as a special case. As stated in Corollary 1 in Ouadah et al., 2020, it yields $\mathbb{E}_{H_0}(V_n) = n^{-1}(n-1)(n-2)p(1-p)$ and $\mathbb{V}\mathrm{ar}_{H_0}(V_n) = n^{-3}2(n-1)(n-2)^2p(1-p)(1+(n-6)p(1-p))$. In order to formulate a feasible test, we make use of Proposition 3.

Proposition 4 *Let $G \in \mathcal{G}_{ER}$ be a homogeneous ER graph $\mathcal{G}(n,p)$ with $p = p(n) \xrightarrow[n\to\infty]{} p_0 \in [0,1)$ such that $np \xrightarrow[n\to\infty]{} \infty$. Then, the degree variance V_n standardized by its estimated moments is asymptotically standard normal. That is, we have*

$$\widehat{R}(V_n) = \frac{V_n - \widehat{\mathbb{E}_{H_0}}(V_n)}{\sqrt{\widehat{\mathbb{V}ar_{H_0}}(V_n)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\widehat{\mathbb{E}_{H_0}}(V_n) = n^{-1}(n-1)(n-2)\widehat{p}(1-\widehat{p})$ and $\widehat{\mathbb{V}ar_{H_0}}(V_n) = n^{-3}2(n-1)(n-2)^2\widehat{p}(1-\widehat{p})(1+(n-6)\widehat{p}(1-\widehat{p}))$. An asymptotic level-$\alpha$ test for testing the null hypothesis H_0 of an underlying homogeneous Erdös-Rényi graph against heterogeneous alternatives H_1 in (4.1) is then given by the decision rule

$$\phi_\alpha(V_n) = \begin{cases} 1, & \text{if } \left|\widehat{R}(V_n)\right| > z_{1-\alpha/2}, \\ 0, & \text{else.} \end{cases}$$

In Section 4.2.2, we already argued that V_n is a graph functional. In order to find its representation following Proposition 2, we consider the centered version $V_n - \mathbb{E}(V_n)$, where $\mathbb{E}(V_n) = n^{-1}(n-1)(n-2)p(1-p)$, and use the Hoeffding decomposition [Hoeffding, 1992; Lee, 2019] as in Ouadah et al., 2020. Then, under the null of a homogeneous ER graph, we get

$$\begin{aligned} V_n - \mathbb{E}(V_n) &= \frac{2(n-2)}{n^2}(1-2p)\sum_{1\le i<j\le n}\widetilde{A}_{ij} + \frac{2(n-4)}{n^2}\sum_{1\le i<j<k\le n}\left(\widetilde{A}_{ij}\widetilde{A}_{ik} + \widetilde{A}_{ij}\widetilde{A}_{jk} + \widetilde{A}_{ik}\widetilde{A}_{jk}\right) \\ &\quad - \frac{8}{n^2}\sum_{1\le i<j<k<l\le n}\left(\widetilde{A}_{ij}\widetilde{A}_{kl} + \widetilde{A}_{ik}\widetilde{A}_{jl} + \widetilde{A}_{il}\widetilde{A}_{jk}\right) \qquad (4.11) \\ &=: V_n^{(A)} + V_n^{(B)} + V_n^{(C)}, \end{aligned}$$

where $\widetilde{A}_{ij} = A_{ij} - p$ are the centered entries of the adjacency matrix. While V_n does not depend on the usually unknown probability p, $\mathbb{E}(V_n)$ does. However, according to Proposition 4, we can use $\widehat{\mathbb{E}}(V_n)$, which is obtained by replacing p by $\widehat{p}$ in $\mathbb{E}(V_n)$, leading to the same asymptotics.

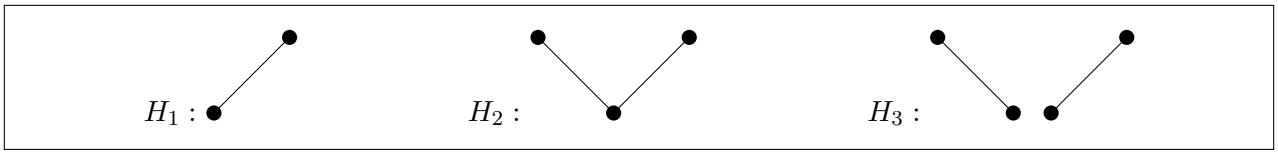


Figure 16: Subgraphs in the decomposition of the degree variance V_n, where $H_1 = P_2$ and $H_2 = P_3$.

The components of the above decomposition are uncorrelated. This simplifies the derivation of the moments of V_n, especially of its variance, leading to

$$\mathbb{V}\mathrm{ar}(V_n) = \mathbb{V}\mathrm{ar}\Big(V_n^{(A)}\Big) + \mathbb{V}\mathrm{ar}\Big(V_n^{(B)}\Big) + \mathbb{V}\mathrm{ar}\Big(V_n^{(C)}\Big).$$

Consequently, the variance $\mathbb{V}\mathrm{ar}(V_n)$ decomposes into three parts $\mathbb{V}\mathrm{ar}(V_n^{(A)})$, $\mathbb{V}\mathrm{ar}(V_n^{(B)})$ and $\mathbb{V}\mathrm{ar}(V_n^{(C)})$ that correspond to the components that are depicted in Figure 16 and can be calculated by

$$\begin{aligned}\mathbb{V}\mathrm{ar}(V_n) \;=\; & \frac{2(n-1)(n-2)^2}{n^3}\,(1-2p)^2\,p(1-p) + \frac{2(n-1)(n-2)(n-4)^2}{n^3}\,\Big(p(1-p)\Big)^2 \\ & + \frac{8(n-1)(n-2)(n-3)}{n^3}\,\Big(p(1-p)\Big)^2.\end{aligned}$$

For the derivation of the limiting normal distribution in Proposition 4, we assumed $np \to \infty$. In view of Theorem 1, the weaker assumption $np^{2/3} \to \infty$ seems to be sufficient. However, a closer inspection of the three parts of $\mathbb{V}\mathrm{ar}(V_n)$ yields that the first on is of order $O(p)$, the second one $O(np^2)$ and the third one $O(p^2)$. Hence, if $np^{2/3} \to \infty$, we have to distinguish three cases: if $np \to \infty$, the second term $V_n^{(B)}$ becomes the leading term. If $np = O(1)$, both the first and the second term, that is $V_n^{(A)}$ and $V_n^{(A)}$ become the leading terms. If $np \to 0$, the first term $V_n^{(A)}$ becomes the leading term. That is, whenever, np does not diverge to ∞, the degree variance V_n does asymptotically depend on the first term which corresponds to $S_n(P2)$ as described below, which is demonstrated to not lead to a useful goodness-of-fit test statistic.

Contrary to the presentations in Ouadah et al., 2020, we argue that it is not $V_n^{(A)}$, but the second part $V_n^{(B)}$, that makes the main contribution to the variance of V_n, if $np \xrightarrow[n\to\infty]{} \infty$. Intuitively, this allocation of the variances makes sense. The information contained in part $V_n^{(A)}$ is the centered number of edges in the observed network. This does not tell us much about the structure of the graph. However, part $V_n^{(B)}$ counts the number of edges that share a common node and are active – it is a rescaled version of the centered count of two-stars P_3 (paths on three vertices with length two). This part of the decomposition contains information on the local structure of the observed network. Part $V_n^{(C)}$ counts the number of pairs of *disjoint* edges that are both active, again an information that does not tell us much about the global or local structure. Hence, it is reasonable that $V_n^{(B)}$ contains most of the information of the network and for that explains most of the variance. The decomposition into $V_n^{(A)}$, $V_n^{(B)}$ and

$V_n^{(C)}$ is the linear combination of centered counts of the substructures $H_1 = P_3$, $H_2 = P_2$ and H_3, illustrated in Figure 16. Thus, alternatively, we can represent (4.11) as

$$\begin{aligned} V_n - \mathbb{E}(V_n) &= \frac{2(n-2)}{n^2}(1-2p)\frac{1}{\mathrm{aut}(H_1)}S_n(H_1) + \frac{2(n-4)}{n^2}\frac{1}{\mathrm{aut}(H_2)}S_n(H_2) - \frac{8}{n^2}\frac{1}{\mathrm{aut}(H_3)}S_n(H_3) \\ &= \frac{(n-2)}{n^2}(1-2p)S_n(H_1) + \frac{(n-4)}{n^2}S_n(H_2) - \frac{1}{n^2}S_n(H_3) \end{aligned}$$

Note that compared to (4.11), we need the prefactors $\frac{1}{\mathrm{aut}(H_i)}$ for $i \in \{1,2,3\}$ because the definition of $S_n(H)$ considers all automorphisms of a given substructure, whereas the sum indices in (4.11) include each node combination only once. Embedded into Proposition 2, as $\mathrm{aut}(H_1) = 2$, $\mathrm{aut}(H_2) = 2$, and $\mathrm{aut}(H_3) = 8$, we have

$$a_n(\emptyset) = \mathbb{E}_{H_0}(V_n),\ a_n(H_1) = \frac{(n-2)}{n^2}(1-2p),\ a_n(H_2) = \frac{(n-4)}{n^2},\ \text{ and }\ a_n(H_3) = -\frac{1}{n^2}.$$

as well as

$$\begin{aligned} &\mathbb{V}\mathrm{ar}(S_n(H_1)) = 2n(n-1)p(1-p),\ \mathbb{V}\mathrm{ar}(S_n(H_2)) = 2n(n-1)(n-2)(p(1-p))^2,\ \text{ and} \\ &\mathbb{V}\mathrm{ar}(S_n(H_3)) = 8n(n-1)(n-2)(n-3)(p(1-p))^2. \end{aligned}$$

Eventually, Theorem 1 can be applied. For this, we need to check if its assumptions are fulfilled. Firstly, a dominating family of connected graphs is required. As $np \to \infty$, both $a_n^2(H_1)\mathbb{V}\mathrm{ar}(S_n(H_1))$ and $a_n^2(H_3)\mathbb{V}\mathrm{ar}(S_n(H_3))$ are of lower order than $a_n^2(H_2)\mathbb{V}\mathrm{ar}(S_n(H_2))$, we set $\mathcal{H}_0 := \{H_2\}$. Then, we can verify that $\mathcal{H}_0$ is a dominating family of connected graphs for V_n by checking the corresponding condition. Indeed, we have

$$\frac{\mathbb{V}\mathrm{ar}(V_n)}{\sum\limits_{H\in\mathcal{H}_\prime} a_n^2(H)\,\mathbb{V}\mathrm{ar}(S_n(H))} = \frac{\frac{2(n-1)(n-2)^2}{n^3}p(1-p)(1+(n-6)p(1-p))}{2\frac{(n-4)^2}{n^4}n(n-1)(n-2)(p(1-p))^2} = \frac{n-2+(n^2-8n+12)}{(n^2-8n+16)} \xrightarrow[n\to\infty]{} 1.$$

Consequently, the variance of V_n is explained by the dominating family $\mathcal{H}_0$. Asymptotically, it is solely driven by part $V_n^{(B)}$ of the Hoeffding decomposition. As a second step, we need to check whether the coefficient

$$b(H_2) = \sup_{n\geq 4} \frac{n^{3/2}p\ \frac{(n-4)}{n^2}}{\sqrt{\frac{2(n-1)(n-2)^2}{n^3}p(1-p)(1+(n-6)p(1-p))}}$$

is finite. Indeed, the finiteness of the supremum above is implied by the convergence

$$\lim_{n\to\infty} \frac{(n^2-8n+16)n^2p}{2(n-1)(n^2-4n+4)(1-p)(1+(n-6)p(1-p))} = \frac{1}{2(1-p_0)^2} < \infty \quad \forall\, p_0 \in [0,1),$$

Furthermore, as the dominating family $\mathcal{H}_0$ is finite, this obviously leads to

$$b^2(H_2)\text{aut}(H_2) = \frac{1}{(1-p_0)^2} < \infty \quad \forall\, p_0 \in [0,1). \tag{4.12}$$

Thus, from (4.11) - (4.12), all conditions of Theorem 1 are fulfilled, yielding the asymptotic normality of $V_n - \mathbb{E}(V_n)$. That is, for $n \to \infty$, we have

$$\frac{V_n - \mathbb{E}(V_n)}{\sqrt{\mathbb{V}\text{ar}(V_n)}} \xrightarrow{d} \mathcal{N}(0,1). \tag{4.13}$$

In particular, for $c_n^2 = (np^2(1-p)^2)^{-1}$, we have $c_n^2 \mathbb{V}\text{ar}(X_n) \to 2$ and $c_n(X_n - \mathbb{E}(X_n)) \xrightarrow{d} \mathcal{N}(0,2)$. As mentioned above, the true underlying parameter p is usually not known in practice. An implementable result for the shown asymptotics is given in Proposition 4 above. Details on the embedding of $\widehat{p}$ in this context can be found in Appendix A.

4.3.2 Goodness-of-fit testing based on centered subgraph counts

In view of the general decomposition (4.4) for arbitrary graph functionals X_n given in Proposition 2, it is natural to make use of centered subgraph counts $S_n(H)$ as building blocks for our unified goodness-of-fit testing approach. However, although the $S_n(H)$'s are graph functionals according to Definition 3, they are not feasible for goodness-of-fit testing as they require the knowledge of the edge probability p. Consequently, the $S_n(H)$'s can not be computed from the adjacency matrix A of a graph G alone. As it is natural to get feasible versions of the $S_n(H)$'s by replacing p by $\widehat{p}$, we define

$$\widehat{S}_n(H) := \sum_{\varphi} \prod_{e \in E(\varphi(H))} \left(\mathbb{1}\{e \in E(G)\} - \widehat{p}\right),$$

Moreover, under the assumptions of Theorem 1 (see also Theorem 6.43 in Janson et al., 2011), it is guaranteed that $S_n(H)$ converges to a standard normal distribution $\mathcal{N}(0,1)$ when divided by the square root of its variance $\mathbb{V}\text{ar}(S_n(H))$. In particular, this holds for any non-null subgraph H. However, according to Example 1, in addition to the null graph $\emptyset$, also P_2 causes problems and is not suitable for goodness-of-fit testing, because it is a too simplistic subgraph structure. Fortunately, as stated in the following proposition, all subgraph structures H that are more complex than P_2 allow to use $\widehat{S}_n(H)$ divided by $\sqrt{\widehat{\mathbb{V}\text{ar}_{H_0}}(S_n(H))}$ instead of $S_n(H)$ divided by $\sqrt{\mathbb{V}\text{ar}_{H_0}(S_n(H))}$ without changing the limiting distribution.

Proposition 5 *Let $G \in \mathcal{G}_{ER}$ be a homogeneous ER graph $\mathcal{G}(n,p)$ with $p = p(n) \xrightarrow[n\to\infty]{} p_0 \in [0,1)$. Further, let H be a non-null connected subgraph with $H \neq P_2$ and assume $np^{r(H)} \xrightarrow[n\to\infty]{} \infty$.*

Then, we have

$$\widehat{S}_n(H) = \sum_{\substack{J \subseteq H \\ J \text{ induced}}} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) \, (p - \widehat{p})^{m(H)-m(J)}, \tag{4.14}$$

where $k(J)$ denotes the number of isolated vertices in J, $\widetilde{J}$ denotes the graph obtained from J after removing all isolated vertices. This leads to

$$\frac{\widehat{S}_n(H)}{\sqrt{\widehat{\mathbb{V}ar}_{H_0}(S_n(H))}} = \frac{\widehat{S}_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(\widehat{p}(1-\widehat{p}))^{m(H)}}} \xrightarrow{d} \mathcal{N}(0,1).$$

We then get an asymptotic level-α test for testing the null hypothesis H_0 of an underlying homogeneous Erdös-Rényi graph against heterogeneous alternatives H_1 in (4.7) *by the decision rule*

$$\phi_\alpha(S_n(H)) := \begin{cases} 1, & \textit{if } \left| \frac{\widehat{S}_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(\widehat{p}(1-\widehat{p}))^{m(H)}}} \right| > z_{1-\alpha/2}, \\ 0, & \textit{else}, \end{cases}$$

That is, according to Proposition 5, using feasible centered subgraph counts $\widehat{S}_n(H)$ instead of infeasible ones $S_n(H)$ and dividing by the square root of the estimated variance of $S_n(H)$ (under H_0) does not change the asymptotic distribution, which is still standard normal $\mathcal{N}(0,1)$. Note that we do not subtract an estimated mean $\widehat{\mathbb{E}_{H_0}}(S_n(H))$ here, because $\mathbb{E}(S_n(H)) = 0$ holds. Instead, we are adjusting $S_n(H)$ itself to get $\widehat{S}_n(H)$, which then can be computed from A alone. This leads to the formulation of asymptotic goodness-of-fit tests for statistics of shape $S_n(H)$.

We illustrate the above result for feasible centered subgraph counts $\widehat{S}_n(H)$ for the special case of a triangle $H = C_3$, i.e., $\widehat{S}_n(C_3)$, in the following example. Raw subgraph counts $T_n(H)$ will be discussed in Section 4.3.3, where we also give an example of the corresponding raw subgraph count $T_n(C_3)$.

Example 4. [Asymptotics for centered counts of triangles] Under the assumptions of Theorem 1, for the infeasible centered subgraph count $S_n(C_3)$, we have

$$\frac{S_n(C_3) - \mathbb{E}_{H_0}(C_3)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(C_3)}} = \frac{S_n(C_3)}{\sqrt{\mathrm{aut}(C_3) n_{n(C_3)} (p(1-p))^{m(C_3)}}} = \frac{S_n(C_3)}{\sqrt{6n(n-1)(n-2)(p(1-p))^3}} \xrightarrow{d} \mathcal{N}(0,1)$$

and, according to Proposition 5, for feasible subgraph counts $\widehat{S}_n(C_3)$, we have the decomposition (see (4.21) in the Appendix)

$$\begin{aligned}\widehat{S}_n(C_3) = &\sum_{\substack{J\subseteq C_3\\ J \text{ induced}}} \left(\prod_{h=n(C_3)-k(J)}^{n(C_3)-1}(n-h)\right) S_n(\widetilde{J})\,(p-\widehat{p})^{m(C_3)-m(J)}\\ =&S_n(C_3) + 3\left(\prod_{h=n(C_3)-0}^{n(C_3)-1}(n-h)\right) S_n(P_3)\,(p-\widehat{p}) + 3\left(\prod_{h=n(C_3)-1}^{n(C_3)-1}(n-h)\right) S_n(P_2)\,(p-\widehat{p})^2\\ &+\left(\prod_{h=0}^{n(C_3)-1}(n-h)\right)(p-\widehat{p})^3\\ =&S_n(C_3) + 3S_n(P_3)\,(p-\widehat{p}) + 3(n-2)S_n(P_2)\,(p-\widehat{p})^2 + n(n-1)(n-2)\,(p-\widehat{p})^3,\end{aligned}$$

Here, we have used that there are also three ways to drop one edge from C_3 leading to the same substructure and there are three ways to drop two edges leading to the same substructure, respectively. When dropping two edges, we end up with a graph consisting of three nodes and one edge such that one node is isolated. Hence, $k(J) = 1$ and $\widetilde{J} = P_2$ in this case, where P_2 consist of two vertices and one edge. Dropping all edges leads to three isolated vertices, i.e. $k(J) = 3$ and $\widetilde{J}$ becomes the null graph $\emptyset$ leading to $S_n(\emptyset) = 1$. When dividing by $\sqrt{\text{aut}(C_3)n_{n(C_3)}(\widehat{p}(1-\widehat{p}))^{m(C_3)}} = \sqrt{6n(n-1)(n-2)(\widehat{p}(1-\widehat{p}))^3}$, we get

$$\frac{3S_n(P_3)\,(p-\widehat{p})}{\sqrt{6n(n-1)(n-2)(\widehat{p}(1-\widehat{p}))^3}} = O_P\left(\frac{\sqrt{n^3(p(1-p))^2}(\frac{\sqrt{p}}{n})}{\sqrt{n^3(p(1-p))^3}}\right) = O_P\left(\frac{1}{n\sqrt{1-p}}\right) = O_P\left(\frac{1}{n}\right) = o_P(1),$$

$$\frac{3(n-2)S_n(P_2)\,(p-\widehat{p})^2}{\sqrt{6n(n-1)(n-2)(\widehat{p}(1-\widehat{p}))^3}} = O_P\left(\frac{n\sqrt{n^2(p(1-p))}(\frac{\sqrt{p}}{n})^2}{\sqrt{n^3(p(1-p))^3}}\right) = O_P\left(\frac{1}{(1-p)^2n^{3/2}}\right) = O_P\left(\frac{1}{n^{3/2}}\right) = o_P(1),$$

$$\frac{n(n-1)(n-2)\,(p-\widehat{p})^3}{\sqrt{6n(n-1)(n-2)(\widehat{p}(1-\widehat{p}))^3}} = O_P\left(\frac{n^3(\frac{\sqrt{p}}{n})^3}{\sqrt{n^3(p(1-p))^3}}\right) = O_P\left(\frac{1}{n^{3/2}(1-p)^{3/2}}\right) = O_P\left(\frac{1}{n^{3/2}}\right) = o_P(1).$$

Altogether, according to Proposition 5, for the feasible centered subgraph count $\widehat{S}_n(C_3)$, we have

$$\frac{\widehat{S}_n(C_3)}{\sqrt{\widehat{\mathbb{V}\text{ar}_{H_0}}(C_3)}} = \frac{\widehat{S}_n(C_3)}{\sqrt{6n(n-1)(n-2)(\widehat{p}(1-\widehat{p}))^3}} \xrightarrow{d} \mathcal{N}(0,1).$$

Hence, $S_n(C_3)$ divided by $\sqrt{\text{aut}(C_3)(n)_3(p(1-p))^3}$ and its counterpart, that is, $\widehat{S}_n(C_3)$ divided by $\sqrt{\text{aut}(C_3)(n)_3(\widehat{p}(1-\widehat{p}))^3}$ converge to the same limiting (standard normal) distribution.

4.3.3 Goodness-of-fit testing based on raw subgraph counts

In contrast to centered subgraph counts $S_n(H)$, which form the building blocks of general graph functionals according to (4.4), but are not feasible as they require the knowledge of p, *raw* subgraph counts $T_n(H)$ also appear to be a natural and, in particular, feasible goodness-of-fit test statistic. Moreover, as the $T_n(H)$'s are graph functionals by construction, under the assumptions of Theorem 1, we have

$$\frac{T_n(H) - \mathbb{E}_{H_0}(T_n(H))}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(T_n(H))}} \xrightarrow{d} \mathcal{N}(0,1).$$

However, both $\mathbb{E}_{H_0}(T_n(H))$ and $\mathbb{V}\mathrm{ar}_{H_0}(T_n(H))$ will generally depend on the unknown parameter p and have to be replaced by suitable estimators $\widehat{\mathbb{E}_{H_0}}(T_n(H))$ and $\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(T_n(H))$ obtained by replacing p by $\widehat{p}$. But, as the $T_n(H)$'s are graph functionals as well, for studying their properties, we can expand it in terms of certain centered subgraph counts using (4.4). Hence, following exactly the approach to decompose $\widehat{S}_n(H)$ in terms of (lower order) centered subgraph counts $S_n(\widetilde{J})$'s leading to (4.14) in Proposition 5, we get the following result for raw subgraph counts.

Proposition 6 *Let $G \in \mathcal{G}_{ER}$ be a homogeneous ER graph $\mathcal{G}(n,p)$ and let H be a non-null connected subgraph. Then, we have*

$$T_n(H) = \frac{1}{\mathrm{aut}(H)} \sum_{\substack{J \subseteq H \\ J \text{ induced}}} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) p^{m(H)-m(J)}, \tag{4.15}$$

where $k(J)$ denotes the number of isolated vertices in J, $\widetilde{J}$ denotes the graph obtained from J after removing all isolated vertices.

For the case of $H = C_3$, we illustrate the consequences of (4.15) for the dominating family of subgraphs in the expansion (4.4) of the graph functional $T_n(C_3)$.

Example 5. [**Asymptotics for raw counts of triangles**] If the assumptions of Theorem 1 hold, for the raw subgraph count $T_n(C_3)$, we have

$$R(T_n(C_3)) = \frac{T_n(C_3) - \mathbb{E}_{H_0}(T_n(C_3))}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(T_n(C_3))}} \xrightarrow{d} \mathcal{N}(0,1).$$

Hence, to check the conditions imposed in Theorem 1, we have to study the decomposition of $T_n(C_3)$ in terms of $S_n(\widetilde{J})$'s. According to Proposition 6, for $H = C_3$, we get

$$\begin{aligned}
T_n(C_3) =& \frac{1}{\mathrm{aut}(C_3)} \sum_{J \subseteq C_3} \left(\prod_{h=n(C_3)-k(J)}^{n(C_3)-1} (n-h) \right) S_n(\widetilde{J}) p^{m(C_3)-m(J)} \\
=& \frac{1}{\mathrm{aut}(C_3)} \left[S_n(J) + 3 \left(\prod_{h=n(C_3)-0}^{n(C_3)-1} (n-h) \right) S_n(P_3) p + 3 \left(\prod_{h=n(C_3)-1}^{n(C_3)-1} (n-h) \right) S_n(P_2) p^2 \right. \\
& \left. + \left(\prod_{h=0}^{n(C_3)-1} (n-h) \right) p^3 \right] \\
=& \frac{1}{\mathrm{aut}(C_3)} \left(S_n(C_3) + 3S_n(P_3)p + 3(n-2)S_n(P_2)p^2 + n(n-1)(n-2)p^3 \right).
\end{aligned}$$

Consequently, using Proposition 1, for the variance of $T_n(C_3)$, we get

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{H_0}(T_n(C_3)) =& \frac{1}{\mathrm{aut}(C_3)^2} \left(\mathbb{V}\mathrm{ar}_{H_0}(S_n(C_3)) + 9\mathbb{V}\mathrm{ar}_{H_0}(S_n(P_3))p^2 + 9(n-2)^2 \mathbb{V}\mathrm{ar}_{H_0}(S_n(P_2))p^4 \right) \\
=& \frac{1}{\mathrm{aut}(C_3)^2} \left(\mathrm{aut}(C_3)(n)_3 (p(1-p))^3 + 9\mathrm{aut}(P_3)(n)_3 (p(1-p))^2 p^2 \right. \\
& \left. + 9(n-2)^2 (n)_2 \mathrm{aut}(P_2) p(1-p) p^4 \right) \\
=& \frac{(n)_3}{\mathrm{aut}(C_3)} p^3 (1-p)^3 + \frac{9\mathrm{aut}(P_3)(n)_3}{\mathrm{aut}(C_3)^2} p^4 (1-p)^2 + \frac{9(n-2)^2 (n)_2 \mathrm{aut}(P_2)}{\mathrm{aut}(C_3)^2} p^5 (1-p) \\
=& \binom{n}{3} p^3 (1-p)^3 + 3\binom{n}{3} p^4 (1-p)^2 + 3(n-2)\binom{n}{3} p^5 (1-p) \\
=& \mathbb{V}\mathrm{ar}_{H_0}^{(1)}(T_n(C_3)) + \mathbb{V}\mathrm{ar}_{H_0}^{(2)}(T_n(C_3)) + \mathbb{V}\mathrm{ar}_{H_0}^{(3)}(T_n(C_3))
\end{aligned}$$

with an obvious notation for $\mathbb{V}\mathrm{ar}_{H_0}^{(i)}(T_n(C_3))$, $i = 1, 2, 3$. Taking a closer look at these three terms, if $p = p(n) \to p_0 \in [0, 1)$, we see that $\mathbb{V}\mathrm{ar}_{H_0}^{(1)}(T_n(C_3)) = O(n^3 p^3)$, $\mathbb{V}\mathrm{ar}_{H_0}^{(2)}(T_n(C_3)) = O(n^3 p^4)$ and $\mathbb{V}\mathrm{ar}_{H_0}^{(3)}(T_n(C_3)) = O(n^4 p^5)$. Hence, there are different cases to distinguish:

(i) If $p = p(n) \to p_0 \in (0, 1)$, we have that $\mathbb{V}\mathrm{ar}_{H_0}^{(3)}(T_n(C_3))$, i.e., $S_n(P_2)$ dominates.

(ii) If $p = p(n) \to 0$, we have that $\mathbb{V}\mathrm{ar}_{H_0}^{(2)}(T_n(C_3)) = o(\mathbb{V}\mathrm{ar}_{H_0}^{(1)}(T_n(C_3)))$, i.e., $S_n(P_3)$ is dominated by $S_n(C_3)$.

(iia) If $p = p(n) \to 0$ such that $np^2 \to 0$, we have $\mathbb{V}\mathrm{ar}_{H_0}^{(3)}(T_n(C_3)) = o(\mathbb{V}\mathrm{ar}_{H_0}^{(1)}(T_n(C_3)))$ and $S_n(C_3)$ dominates.

(iib) If $p = p(n) \to 0$ such that $np^2 \to K$ for some constant $K > 0$, we have that $\mathbb{V}\mathrm{ar}_{H_0}^{(3)}(T_n(C_3))$ and $\mathbb{V}\mathrm{ar}_{H_0}^{(1)}(T_n(C_3))$ are of the same order and both $S_n(C_3)$ and $S_n(P_2)$ dominate.

(iic) If $p = p(n) \to 0$ such that $np^2 \to \infty$, we have that $\mathbb{V}\mathrm{ar}_{H_0}^{(1)}(T_n(C_3)) = o(\mathbb{V}\mathrm{ar}_{H_0}^{(3)}(T_n(C_3)))$ and $S_n(P_2)$ dominates.

Note that the condition $np^{r(H)} \to \infty$ as $n \to \infty$ imposed in Theorem 1 for the dominating family of subgraphs $H \in \mathcal{H}$ is trivially fulfilled in cases (i), (iib) and (iic), while case (iia) requires also $np^{r(C_3)} = np \to \infty$. Moreover, in view of Example 1, cases (i) and (iic) lead to a goodness-of-fit test dominated by P_2, which is *not* suitable at all to detect deviations from the H_0 of a homogeneous Erdös-Rényi graph. Case (iib) is a mixed case that makes use of P_2 and C_3. Hence, (iib) will be not efficient, while case (iia) has the largest potential to detect deviations from H_0.

However, at this point, it is important to note that $R(T_n(C_3))$ above defines an *infeasible* test statistic as it requires the knowledge of p, because both $\mathbb{E}_{H_0}(T_n(C_3)) = \binom{n}{3}p^3$ and $\mathbb{V}\mathrm{ar}_{H_0}(T_n(C_3))$ as derived above are functions of p. That is, according to our construction principle, instead, we have to replace $\mathbb{E}_{H_0}(T_n(C_3))$ and $\mathbb{V}\mathrm{ar}_{H_0}(T_n(C_3))$ by $\widehat{\mathbb{E}_{H_0}}(T_n(C_3))$ and $\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(T_n(C_3))$, respectively.

Example 5 continued. [**Asymptotics for raw counts of triangles**] Based on the raw subgraph count $T_n(C_3)$, a feasible goodness-of test statistic is obtained by

$$\widehat{R}(T_n(C_3)) = \frac{T_n(C_3) - \widehat{\mathbb{E}_{H_0}}(T_n(C_3))}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(T_n(C_3))}} = \frac{T_n(C_3) - \binom{n}{3}\widehat{p}^3}{\sqrt{\binom{n}{3}\widehat{p}^3(1-\widehat{p})^3 + 3\binom{n}{3}\widehat{p}^4(1-\widehat{p})^2 + 3(n-2)\binom{n}{3}\widehat{p}^5(1-\widehat{p})}}.$$

Let's have a closer look at the numerator. First, using (4.10), we can re-write $\binom{n}{3}\widehat{p}^3$ to get

$$\begin{aligned}\binom{n}{3}\widehat{p}^3 =& \binom{n}{3}\left(\frac{1}{2\binom{n}{2}}S_n(P_2) + p\right)^3 = \binom{n}{3}\left(\left(\frac{1}{2\binom{n}{2}}\right)^3 S_n^3(P_2) + 3\left(\frac{1}{2\binom{n}{2}}\right)^2 S_n^2(P_2)p + 3\frac{1}{2\binom{n}{2}}S_n(P_2)p^2 + p^3\right)\\ =& \frac{\binom{n}{3}}{\left(2\binom{n}{2}\right)^3}S_n^3(P_2) + \frac{3\binom{n}{3}}{\left(2\binom{n}{2}\right)^2}S_n^2(P_2)p + \frac{3\binom{n}{3}}{2\binom{n}{2}}S_n(P_2)p^2 + \binom{n}{3}p^3.\end{aligned}$$

Subtracting the last right-hand side from $T_n(C_3)$ and plugging-in the decomposition for $T_n(C_3)$, leads to

$$\begin{aligned}T_n(C_3) - \widehat{\mathbb{E}_{H_0}}(T_n(C_3)) =& \frac{1}{\mathrm{aut}(C_3)}\left(S_n(C_3) + 3S_n(P_3)p + 3(n-2)S_n(P_2)p^2 + n(n-1)(n-2)p^3\right)\\ &- \left(\frac{\binom{n}{3}}{\left(2\binom{n}{2}\right)^3}S_n^3(P_2) + \frac{3\binom{n}{3}}{\left(2\binom{n}{2}\right)^2}S_n^2(P_2)p + \frac{3\binom{n}{3}}{2\binom{n}{2}}S_n(P_2)p^2 + \binom{n}{3}p^3\right)\\ =& \frac{1}{\mathrm{aut}(C_3)}\left(S_n(C_3) + 3S_n(P_3)p\right) - \left(\frac{\binom{n}{3}}{\left(2\binom{n}{2}\right)^3}S_n^3(P_2) + \frac{3\binom{n}{3}}{\left(2\binom{n}{2}\right)^2}S_n^2(P_2)p\right).\end{aligned}$$

As the latter two terms on the last right-hand are of order $p^{3/2}$ and np^2, which are both slower than the orders of the first two terms, which are $(np)^{3/2}$ and $n^{3/2}p^2$. Hence, the last expression is of leading order $(np)^{3/2}$. Altogether, picking-up the different cases introduced in Example 5,

we get

$$\frac{T_n(C_3) - \widehat{E_{H_0}}(T_n(C_3))}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(T_n(C_3))}} = \frac{T_n(C_3) - \binom{n}{3}\widehat{p}^3}{\sqrt{\binom{n}{3}\widehat{p}^3(1-\widehat{p})^3 + 3\binom{n}{3}\widehat{p}^4(1-\widehat{p})^2 + 3(n-2)\binom{n}{3}\widehat{p}^5(1-\widehat{p})}} \xrightarrow{P} 0,$$

for (i) and (iic), because $(np)^{3/2}/n^2p^{5/2} = \sqrt{\frac{1}{np^2}} \to 0$ in both cases. For (iib), we get

$$\frac{T_n(C_3) - \widehat{E_{H_0}}(T_n(C_3))}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(T_n(C_3))}} = \frac{T_n(C_3) - \binom{n}{3}\widehat{p}^3}{\sqrt{\binom{n}{3}\widehat{p}^3(1-\widehat{p})^3 + 3\binom{n}{3}\widehat{p}^4(1-\widehat{p})^2 + 3(n-2)\binom{n}{3}\widehat{p}^5(1-\widehat{p})}} \xrightarrow{d} \mathcal{N}(0,\sigma^2),$$

where

$$\sigma^2 := \lim_{n\to\infty} \frac{\binom{n}{3}p^3(1-p)^3 + 3\binom{n}{3}p^4(1-p)^2}{\binom{n}{3}p^3(1-p)^3 + 3\binom{n}{3}p^4(1-p)^2 + 3(n-2)\binom{n}{3}p^5(1-p)} = \frac{1}{1+3K},$$

where $K = \lim_{n\to\infty} np^2$. Finally, for (iia), we get

$$\frac{T_n(C_3) - \widehat{E_{H_0}}(T_n(C_3))}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(T_n(C_3))}} = \frac{T_n(C_3) - \binom{n}{3}\widehat{p}^3}{\sqrt{\binom{n}{3}\widehat{p}^3(1-\widehat{p})^3 + 3\binom{n}{3}\widehat{p}^4(1-\widehat{p})^2 + 3(n-2)\binom{n}{3}\widehat{p}^5(1-\widehat{p})}} \xrightarrow{d} \mathcal{N}(0,1).$$

Hence, using standard normal quantiles for implementing a goodness-of-fit test based on $T_n(C_3)$ will be asymptotically valid only for the case of $p = p(n) \to 0$ such that $np^2 \to 0$. That is, in the case of (iia), where $S_n(C_3)$ dominates. This observation coincides with the findings in Example 6.50 in Janson et al., 2011.

In view of the examples discussed above and assumptions (A) and (B) imposed in Proposition 3, the construction principle of goodness-of-fit tests is flexible and allows for a unified treatment based on the theory on graph functionals gathered in Section 4.2.2. Nevertheless, one has to be careful as condition (B) may be hurt such that $\widehat{R}(X_n)$ turns out to have a different asymptotic behavior than $R(X_n)$, which may lead to tests of incorrect size or tests that have (asymptotically) no power.

4.3.4 Power Analysis for SBM alternatives

For any given graph functional X_n, the feasible test $\widehat{R}(X_n)$ (that makes use of standard normal quantiles) requires conditions (A) and (B) in (4.9) to hold under H_0. The same test will have non-trivial power for (certain) alternatives under H_1, if

$$\left| \frac{\mathbb{E}_{H_1}(X_n) - \widehat{\mathbb{E}_{H_0}}(X_n)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n)}} \right| \xrightarrow{P} \infty \quad \text{as} \quad n \to \infty$$

holds. As the class of graph functionals is very broad, there are a lot of different possible test statistics to choose from. Obviously, each possible graph functional might be sensitive to different network information and thus might be more or less suitable for a particular data scenario. On the one hand, this leads to a large flexibility in constructing goodness-of-fit tests for various application fields. On the other hand, the derivation of general consistency results for a goodness-of-fit test $\widehat{R}(X_n)$ for arbitrary graph functionals X_n and general alternatives under H_1 will not be possible. However, the question of which graph functional to use in which situation is pretty crucial in practice. Because general statements are obviously very difficult to derive, we provide a concrete theory-driven analysis for centered as well as raw subgraph counts that rely on triangles, denoted as C_3, and two-stars, denoted as P_3, below. Further examples are then investigated in a simulation study in Section 4.5.

For now, to generate stochastic networks under the alternative, we consider the popular case of Stochastic Block Models (SBM) that particularly plays an important role in modeling social network behaviors and patterns. In general, it assumes that the set of n nodes can be divided into K blocks. Typically, nodes of the same block have a higher probability to share an edge than nodes of different blocks. Hence, SBMs are low-parametrized special cases of $\mathcal{G}_{HER}(n)$ models. Throughout this subsection, to make explicit calculations and derivations tractable, we consider only SBMs with $K = 2$ blocks, equal block sizes (i.e. $n/2$) and equal intra-group probabilities p_{intra} for both blocks. The edge probability between the groups are denoted by p_{inter} such that $p_{\text{intra}} \geq p_{\text{inter}}$. For $p_{\text{intra}} = p_{\text{inter}}$, the SBM becomes a $\mathcal{G}_{ER}(n)$ model. To rule out detection just based on the total number of edges, we generate models from both model classes such that their mean connectivity parameters p_{mean} (which is allowed to depend on n) coincide. Note that, obviously, $p_{\text{mean}} = p$ for the $\mathcal{G}_{ER}(n)$ model class. Let us first focus on the performance of C_3.

Theorem 2 (Power of $S_n(C_3)$ and $T_n(C_3)$ under fixed and local alternatives) *Suppose the network G is either generated by a homogeneous ER model $\mathcal{G}(n,p)$ with edge probability p or by a stochastic block model (SBM) with $K = 2$ blocks, equal block sizes $n/2$, equal intra-group probability p_{intra} for both blocks such that*

$$p_{\text{intra}} - p_{\text{inter}} = \epsilon \quad \text{and} \quad p_{\text{mean}} = p,$$

where

$$p_{\text{mean}} := \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} p_{ij} = \frac{2\binom{n/2}{2} p_{\text{intra}} + \left(\frac{n}{2}\right)^2 p_{\text{inter}}}{\binom{n}{2}} = \frac{n-2}{2n-2} p_{\text{intra}} + \frac{n}{2n-2} p_{\text{inter}}$$

denotes the mean edge probability (mean connectivity) p_{mean} of the SBM. Further, let $\mathbb{E}_{\text{ER}}(\cdot) = \mathbb{E}_{H_0}(\cdot)$ and $\mathbb{V}ar_{\text{ER}}(\cdot) = \mathbb{V}ar_{H_0}(\cdot)$ denote expectation and variance, respectively, under the ER model and let $\mathbb{E}_{\text{SBM}}(\cdot) = \mathbb{E}_{H_1}(\cdot)$ denote the expectation under the SBM. Then, for the centered

subgraph count $S_n(C_3)$ and the raw subgraph count $T_n(C_3)$, respectively, for $n \to \infty$, the following holds:

(i) ***Fixed alternatives:*** *If $p \to p_0 \in [0,1)$ with $np \to \infty$ and $\epsilon \sim p$, we have*

$$(a) \quad \frac{\mathbb{E}_{\text{SBM}}(S_n(C_3)) - \mathbb{E}_{\text{ER}}(S_n(C_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(S_n(C_3))}} \to \begin{cases} O(n^{3/2}), & \text{if } p \to p_0 \in (0,1) \\ O((np)^{3/2}), & \text{if } p \to 0, \end{cases}$$

and

$$(b) \quad \frac{\mathbb{E}_{\text{SBM}}(T_n(C_3)) - \mathbb{E}_{\text{ER}}(T_n(C_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(T_n(C_3))}} \to \begin{cases} O(n), & \text{if } p \to p_0 \in (0,1) \\ O(n\sqrt{p}), & \text{if } p \to 0,\ np^2 \to \infty \\ O((np)^{3/2}), & \text{if } p \to 0,\ np^2 \to K \geq 0 \end{cases}$$

Hence, if $np \to \infty$, we get divergence to $+\infty$ in all cases.

(ii) ***Local alternatives:*** *If $p \to p_0 \in [0,1)$ with $np \to \infty$ and $\epsilon = o(p)$, we have*

$$(a) \quad \frac{\mathbb{E}_{\text{SBM}}(S_n(C_3)) - \mathbb{E}_{\text{ER}}(S_n(C_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(S_n(C_3))}} = \begin{cases} O(n^{3/2}\epsilon^3), & \text{if } p \to p_0 \in (0,1) \\ O(n^{3/2}\epsilon^3 p^{-3/2}), & \text{if } p \to 0, \end{cases}$$

and

$$(b) \quad \frac{\mathbb{E}_{\text{SBM}}(T_n(C_3)) - \mathbb{E}_{\text{ER}}(T_n(C_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(T_n(C_3))}} = \begin{cases} O(n\epsilon^3 p^{-5/2}), & \text{if } n\epsilon \to \infty,\ np^2 \to \infty \\ O(\epsilon p^{-5/2}), & \text{if } n\epsilon \to K_\epsilon \geq 0,\ np^2 \to \infty \\ O(n^{3/2}\epsilon^3 p^{-3/2}), & \text{if } n\epsilon \to \infty,\ np^2 \to K \geq 0 \\ O(n^{1/2}\epsilon p^{-3/2}), & \text{if } n\epsilon \to K_\epsilon \geq 0,\ np^2 \to K \geq 0 \end{cases}.$$

Whenever, in one of the above cases, $\epsilon \to 0$ such that the corresponding expression in $O(\cdot)$ notation is bounded, local power is obtained. Whenever $\epsilon \to 0$ at any faster rate, we get $o(1)$.

In the context of our work, we are more interested in statements (a) about the behavior of the centered subgraph count $S_n(C_3)$. However, for both $S_n(C_3)$ and $T_n(C_3)$, the theorem gives conditions such that C_3 is sensitive to distinguishing between the investigated models. Similar to these derivations, we can derive a corresponding result for P_3.

Theorem 3 (Power of $S_n(P_3)$ and $T_n(P_3)$ under fixed and local alternatives) *With the same setup as for Theorem 2, the following holds for $n \to \infty$:*

(i) ***Fixed alternatives:*** *If $p \to p_0 \in [0,1)$ with $np^{2/3} \to \infty$ and $\epsilon \sim p$, we have*

$$(a) \quad \frac{\mathbb{E}_{\text{SBM}}(S_n(P_3)) - \mathbb{E}_{\text{ER}}(S_n(P_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(S_n(P_3))}} \to \begin{cases} O(n^{1/2}), & \text{if } p \to p_0 \in (0,1) \\ O(n^{1/2}p), & \text{if } p \to 0 \end{cases}$$

and

$$(b) \quad \frac{\mathbb{E}_{\text{SBM}}(T_n(P_3)) - \mathbb{E}_{\text{ER}}(T_n(P_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(T_n(P_3))}} \to \begin{cases} O(1), & \text{if } p \to p_0 \in (0,1) \\ O(\sqrt{p}), & \text{if } p \to 0,\ np \to \infty \\ O(\sqrt{n}p), & \text{if } p \to 0,\ np \to K \geq 0 \end{cases}$$

Hence, if $np^2 \to \infty$, we get divergence to $+\infty$ for $S_n(P_3)$, but not for $T_n(P_3)$.

(ii) ***Local alternatives:*** *If $p \to p_0 \in [0,1)$ with $np^{2/3} \to \infty$ and $\epsilon = o(p)$, we have*

$$(a) \quad \frac{\mathbb{E}_{\text{SBM}}(S_n(P_3)) - \mathbb{E}_{\text{ER}}(S_n(P_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(S_n(P_3))}} = \begin{cases} O(n^{1/2}\epsilon^2), & \text{if } p \to p_0 \in (0,1) \\ O(n^{1/2}\epsilon^2 p^{-1}), & \text{if } p \to 0 \end{cases}$$

and

$$(b) \quad \frac{\mathbb{E}_{\text{SBM}}(T_n(P_3)) - \mathbb{E}_{\text{ER}}(T_n(P_3))}{\sqrt{\mathbb{V}ar_{\text{ER}}(T_n(P_3))}} = \begin{cases} O(\epsilon), & \text{if } p \to p_0 \in (0,1) \\ O(\epsilon p^{-1/2}), & \text{if } p \to 0,\ np \to \infty \\ O(n^{1/2}\epsilon), & \text{if } p \to 0,\ np \to K \geq 0 \end{cases} .$$

Hence, if $n\epsilon^4 = O(1)$ for the case $p \to p_0 \in (0,1)$ or if $n\epsilon^4 p^{-2} = O(1)$ for $p \to 0$, local power is obtained for $S_n(P_3)$. Whenever $\epsilon \to 0$ at any faster rate, we get $o(1)$. For $T_n(P_3)$, we do not have power.

Considering the rates, the triangle structure is more sensitive to detecting these types of SBMs. Intuitively, this makes sense: The conditions for triangles are more restrictive than for two-stars, since all links between the three involved nodes have to exist. Obviously, this is easier to achieve in the dense blocks of a SBM than in an ER-graph due to $p_{\text{intra}} > p_{\text{mean}}$. These results are confirmed by the simulation study in Section 4.5, where we also evaluate the behaviour for more flexible SBM setups with different parameters.

4.4 Bootstrap theory for graph functionals

While knowledge of the presented limiting distributions for the class of graph functionals generally allows the construction and implementation of a powerful testing procedure, such

asymptotic tests might have issues when it comes to small sample sizes, i.e. small network sizes n. Additionally, the derivation of the centered subgraph representation following Proposition 2 is not trivial and can be tedious. In order to provide a more feasible solution and to potentially profit from finite sample improvements, we propose to use parametric bootstrapping to approximate the distribution of the test under the null.

4.4.1 Bootstrap Scheme

Having observed a simple random graph G, the bootstrap algorithm to estimate the distribution of a graph functional X_n under the null of an ER-graph is defined as follows:

Step 1. Estimate the connection probability p by $\widehat{p} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} A_{ij}$.

Step 2. Conditional on $\boldsymbol{A} = (A_{ij})_{1 \le i,j \le n}$, generate an ER graph G^* by drawing a symmetric adjacency matrix $\boldsymbol{A}^* = (A^*_{ij})_{1 \le i,j \le n}$ from $\mathcal{G}(n, \widehat{p})$. That is, conditional on $\boldsymbol{A} = (A_{ij})_{1 \le i,j \le n}$, we have $A^*_{ij} \sim \mathrm{Bin}(1, \widehat{p})$ for $1 \le i < j \le n$ with $A^*_{ji} := A^*_{ij}$, and $A^*_{ii} := 0$ for all i.

Step 3. Calculate the bootstrap graph functional $X^*_n = X_n(G^*)$.

Step 4. Repeat Steps 2 and 3 B times, where B is large, to obtain $X^{*(b)}_n$, $b = 1, ..., B$.

Step 5. Approximate the distribution of graph functional X_n by the empirical distribution of the bootstrap graph functionals $X^{*(1)}_n, \ldots, X^{*(B)}_n$ (percentile bootstrap) or, alternatively, the distribution of the centered graph functional $X_n - \mathbb{E}(X_n)$ by the empirical distribution of the centered bootstrap graph functionals $X^{*(1)}_n - \mathbb{E}^*(X^*_n), \ldots, X^{*(B)}_n - \mathbb{E}^*(X^*_n)$ (Hall bootstrap), where $\mathbb{E}^*(\cdot)$ denotes the bootstrap expectation conditional on the original network.

4.4.2 Bootstrap Theory

In the following theorem, we provide asymptotic theory for the bootstrap procedure and prove its consistency in the framework of Theorem 1 under the alternative (including the null).

Theorem 4 *Suppose $G \in \mathcal{G}_{\mathrm{HER}}$ with mean connectivity $p_{mean} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} p_{ij}$ such that $p_{mean} = p_{mean}(n) \xrightarrow[n \to \infty]{} p_0 \in [0, 1)$ with $\sup_{i,j} p_{ij} = o(1)$ if $p_0 = 0$ and $\sup_{n;i,j} p_{ij} < 1$ and $\inf_{n;i,j} p_{ij} > 0$ if $p_0 \in (0, 1)$. Further, let X_n be a graph functional $X_n = X_n(G)$ computed from G. Suppose that, under the null hypothesis of a homogeneous ER graph, i.e. when $G \in \mathcal{G}_{\mathrm{ER}} \subsetneq \mathcal{G}_{\mathrm{HER}}$, the graph functional X_n has an expansion $X_n = \sum_{H \in \mathcal{H}} a_n(H) S_n(H)$ with $a_n(H) = a_n(H, p)$ that is dominated by a family of connected graphs $\mathcal{H}_0$. For all $H \in \mathcal{H}_0$, we*

suppose that $np_{\text{mean}}^{r(H)} \xrightarrow[n\to\infty]{} \infty$ *with* $r(H) = \max_{J\subseteq H} d(J) \geq 1$, *where* $d(J) = m(J)/n(J)$ *denotes the density of* J, *the coefficients*

$$b_{mean}(H) = \sup_n \frac{n^{n(H)/2} p_{\text{mean}}^{m(H)/2} a_n(H)}{\sqrt{\mathbb{V}ar(X_n)}} \tag{4.16}$$

are finite and satisfy

$$\sum_{H\in\mathcal{H}_0} b^2(H)|\text{aut}(H)| < \infty. \tag{4.17}$$

Further, let $X_n^* = X_n(G^*)$ *denote the bootstrap version of the graph functional* X_n *of* G^* *that is generated according to the bootstrap scheme in Section 4.4.1. Then, we have* $X_n^* = \sum_{H\in\mathcal{H}} a_n^*(H) S_n^*(H)$ *with* $a_n^*(H) = a_n(H, \widehat{p})$ *and*

$$S_n^*(H) = \sum_{\varphi} \prod_{e\in E(\varphi(H))} \Big(\mathbb{1}\{e \in E(G^*)\} - \mathbb{P}^*(e \in E(G^*))\Big) = \sum_{\varphi} \prod_{e\in E(\varphi(H))} \Big(\mathbb{1}\{e \in E(G^*)\} - \widehat{p}\Big),$$

and, as $n \to \infty$, *it holds*

$$R(X_n^*) := \frac{X_n^* - \mathbb{E}^*(X_n^*)}{\sqrt{\mathbb{V}ar^*(X_n^*)}} \xrightarrow{d} \mathcal{N}(0,1) \quad \textit{in probability.}$$

Additionally, with $(c_n)_{n\in\mathbb{N}}$ *and* V^2 *as defined in Theorem 1, we have* $c_n^2 \mathbb{V}ar^*(X_n^*) \to V^2 > 0$ *in probability.*

As a direct consequence, we get bootstrap consistency for the bootstrap procedure from Section 4.4.1 in the following sense.

Corollary 1 (Bootstrap consistency) *Under the assumptions of Theorems 1 and 4, we have*

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}_{H_0}\left(\frac{X_n - \mathbb{E}_{H_0}(X_n)}{\sqrt{\mathbb{V}ar_{H_0}(X_n)}} \leq x \right) - \mathbb{P}^*\left(\frac{X_n^* - \mathbb{E}^*(X_n^*)}{\sqrt{\mathbb{V}ar^*(X_n^*)}} \leq x \right) \right| \to 0 \quad \textit{in probability,} \tag{4.18}$$

where $\mathbb{P}_{H_0}$ *denotes the probability under the null hypothesis of a homogeneous ER graph and* $\mathbb{P}^*$ *denotes the bootstrap probability measure induced by the bootstrap scheme from Section 4 (conditional on* G*). Additionally, we have*

$$c_n^2 \left(\mathbb{V}ar_{H_0}(X_n) - \mathbb{V}ar^*(X_n^*)\right) \to 0 \quad \textit{in probability} \tag{4.19}$$

leading to

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}_{H_0}\left(c_n(X_n - \mathbb{E}_{H_0}(X_n)) \leq x\right) - \mathbb{P}^*\left(c_n(X_n^* - \mathbb{E}^*(X_n^*)) \leq x\right) \right| \to 0 \quad \textit{in probability,} \tag{4.20}$$

The bootstrap consistency results for the parametric bootstrap procedure in (4.18) and (4.20) enable the construction of goodness-of-fit tests based on graph functionals not only in an asymptotic, but also in a bootstrap manner. This has the great advantage that we do not rely anymore on finding the explicit centered subgraph representation of a graph functional as in Proposition 2 leading to a handily applicable class of goodness-of-fit procedures. In particular, if conditions (A) and (B) in (4.9) hold, an asymptotically valid bootstrap level-α test for testing the null hypothesis H_0 of an underlying homogeneous Erdös-Rényi graph against heterogeneous alternatives H_1 in (4.1) is given by the decision rule

$$\phi_\alpha^*(X_n) = \begin{cases} 1, & \text{if } \widehat{R}(X_n) < z_{\alpha/2}^* \quad \text{or} \quad \widehat{R}(X_n) > z_{1-\alpha/2}^*, \\ 0, & \text{else,} \end{cases}$$

where $z_{\alpha/2}^*$ and $z_{1-\alpha/2}^*$ are the $\alpha/2$ and the $(1-\alpha/2)$ quantiles of the empirical distribution of the resampled statistic $\widehat{R}(X_n^*) = (X_n^* - \widehat{\mathbb{E}}^*(X_n^*))/\sqrt{\widehat{\mathbb{V}\mathrm{ar}}^*(X_n^*)}$. Due to (4.19), an alternative asymptotically valid bootstrap level-α test is given by

$$\widetilde{\phi}_\alpha^*(X_n) = \begin{cases} 1, & \text{if } X_n - \widehat{\mathbb{E}_{H_0}}(X_n) < \widetilde{z}_{\alpha/2}^* \quad \text{or} \quad X_n - \widehat{\mathbb{E}_{H_0}}(X_n) > \widetilde{z}_{1-\alpha/2}^* \\ 0, & \text{else,} \end{cases}$$

where $\widetilde{z}_{\alpha/2}^*$ and $\widetilde{z}_{1-\alpha/2}^*$ are the $\alpha/2$ and the $(1-\alpha/2)$ quantile of the empirical distribution of the resampled statistic $X_n^* - \widehat{\mathbb{E}}^*(X_n^*)$. Precisely, this general form now enables us to formulate suitable bootstrap versions of the asymptotic tests presented in Section 4.3

Example 2 continued. [**Degree Variance**] A bootstrap version of the asymptotic level$-\alpha$ test using the degree variance V_n is given by

$$\phi_\alpha^*(V_n) = \begin{cases} 1, & \text{if } \frac{V_n - \widehat{\mathbb{E}_{H_0}}(V_n)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(V_n)}} < z_{\alpha/2}^* \quad \text{or} \quad \frac{V_n - \widehat{\mathbb{E}_{H_0}}(V_n)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(V_n)}} > z_{1-\alpha/2}^*, \\ 0, & \text{else,} \end{cases}$$

where $z_{\alpha/2}^*$ and $z_{1-\alpha/2}^*$ are the $\alpha/2$ and the $(1-\alpha/2)$ quantile of the empirical distribution of the resampled statistic $\frac{V_n^* - \widehat{\mathbb{E}}^*(V_n^*)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}^*(V_n^*)}}$. An alternative bootstrap test of asymptotic level$-\alpha$ is given by

$$\widetilde{\phi}_\alpha^*(V_n) = \begin{cases} 1, & \text{if } V_n - \widehat{\mathbb{E}_{H_0}}(V_n) < z_{\alpha/2}^* \quad \text{or} \quad V_n - \widehat{\mathbb{E}_{H_0}}(V_n) > z_{1-\alpha/2}^* \\ 0, & \text{else,} \end{cases}$$

where $z_{\alpha/2}^*$ and $z_{1-\alpha/2}^*$ are the $\alpha/2$ and the $(1-\alpha/2)$ quantile of the empirical distribution of the resampled statistic $V_n^* - \widehat{\mathbb{E}}^*(V_n^*)$.

Example 3 continued. [**Eigenvector centrality**] As long as the conditions of Theorem 4 as well as (A) and (B) from Proposition 3 are satisfied, we may use any graph functional for constructing a goodness-of-fit without deriving its centered subgraph count representation according to Proposition 2. This is particularly helpful for those graph functionals for which finding this representation may be tedious. This can be the case for quite complicated network metrics such as centrality metrics. These are popularly used in practice, but to the best of our knowledge there does not exist (yet) a well established asymptotic theory. This is why the bootstrap procedure can be beneficial as it allows us to construct novel tests with statistics that have not been used before. A bootstrap test of asymptotic level$-\alpha$ for the average eigenvector centrality E_n is given by

$$\widetilde{\phi}^*_\alpha(E_n) = \begin{cases} 1, & \text{if } E_n - \widehat{\mathbb{E}_{H_0}}(E_n) < z^*_{\alpha/2} \quad \text{or} \quad E_n - \widehat{\mathbb{E}_{H_0}}(E_n) > z^*_{1-\alpha/2} \\ 0, & \text{else,} \end{cases}$$

where $z^*_{\alpha/2}$ and $z^*_{1-\alpha/2}$ are the $\alpha/2$ and the $(1-\alpha/2)$ quantile of the empirical distribution of the resampled statistic $E^*_n - \widehat{\mathbb{E}}^*(E^*_n)$. To the best of our knowledge there does not exist a closed form of $\mathbb{E}_{H_0}(E_n)$ in the literature. Consequently, there is also no closed form of the plug-in version $\widehat{\mathbb{E}_{H_0}}(E_n)$. This is why we use a double bootstrap approach (see e.g. Efron, 1992) in order to estimate $\widehat{\mathbb{E}_{H_0}}(E_n)$.

4.5 Simulation Study

To underline our findings and to gain further evidence for the proposed class of tests, we execute a simulation study for different parameter setups and various alternatives. We analyze the power of the tests and illustrate performance differences depending on the chosen graph functional.

4.5.1 General Setting

We investigate four different graph functionals for the construction of the test statistics. Namely, these are the degree variance statistic V_n, the average eigenvector centrality E_n, centered triangle counts $S_n(C_3)$ and centered two-star counts $S_n(P_3)$. We analyze their performance for the asymptotic version and for the bootstrap version of the testing procedure. Note that for E_n, we only consider the bootstrap version. To investigate the power of these test statistics, we need to generate networks from the alternative, i.e from the $\mathcal{G}_{\text{HER}}(n)$ model. In order to be able to detect the alternative, the generated networks have to deviate from the homogeneous null model by an increasing amount with increasing heterogeneity. To generate random graphs under the alternative, we set a mean connectivity p_{mean} such that the null model is a $\mathcal{G}(n, p_{\text{mean}})$ model.

This is to ensure that the power of the tests is not due to differences in the mean connectivity (thus differing degrees and mean degrees), but actually due to the rising heterogeneity of the connection probabilities. Hence, we make the requirement:

$$\binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} p_{ij} = p_{\text{mean}}.$$

This preserves the mean connectivity over all scenarios and settings. Although this limits the possible alternatives, it is not a big constraint. Having an observed network G, we can directly calculate its mean connectivity p_{mean}. In order to study the performances in various network sizes and density conditions, we use different values of $n = \{16, 32, 64, 128, 256, 512\}$ and $p_{\text{mean}} = p_{\text{mean}}(n) = \left\{\frac{\log n}{n}, \frac{1}{\sqrt{n}}, \frac{\log n}{\sqrt{n}}\right\}$. Note that for these combinations of n and p_{mean} we have $\frac{\log n}{n} < \frac{1}{\sqrt{n}} < \frac{\log n}{\sqrt{n}}$. For $n = \{256, 512\}$, we will only report the performance of the asymptotic tests as the computation of the power curves of the bootstrap procedure is quite expensive. As we will see, the asymptotic results stand representative for the bootstrap results as they coincide for large n. As the $\mathcal{G}_{\text{HER}}(n)$ model class is very broad, we limit the simulation study to a few relevant scenarios for the alternative that serve as representatives for different possible ways the matrix $\boldsymbol{P}$ can be set up. These include the popular use case of SBMs and covariate models. For the SBMs, we use two different setup versions including the usage of two blocks with equal block sizes as in our analysis of Section 4.3.4 and the extension to three blocks with random block sizes and varying intra-group probabilities. The exact setup of each alternative will be explained in the corresponding subsections below. We calculate the power of the tests by the application of 1000 replications for each setting. For the bootstrap version, we use $B = 700$ bootstrap replications. We set the test level as $\alpha = 0.05$ for all tests.

4.5.2 Performances

With the consideration of all described parameter variations, we investigate the power of 72 different scenarios. The main results and important insights are presented below.

Stochastic block models

As mentioned, we construct two different versions of SBMs. The first one is characterized by two blocks of equal block sizes and is constructed as follows. We assign weights to the nodes with

$$\boldsymbol{w} = (w_1, ..., w_n) = \Big(\underbrace{-\frac{1}{2}, ..., -\frac{1}{2}}_{n/2 \text{ times}}, \underbrace{\frac{1}{2}, ..., \frac{1}{2}}_{n/2 \text{ times}} \Big).$$

By premultiplying the weights with a factor $\lambda > 0$, we can achieve different degrees of heterogeneity. Concretely, the weights are transformed into connection probabilities p_{ij} using a logit link. We set

$$p_{ij}(a, \lambda) = \begin{cases} \frac{\exp(a+\lambda^2 w_i w_j)}{1+\exp(a+\lambda^2 w_i w_j)}, & \text{for } i \neq j \\ 0 & \text{else.} \end{cases}$$

The constant a is set to preserve the mean connectivity p_{mean} through the different heterogeneity levels. This is achieved by (numerically) minimizing the function

$$f(a) = \left| p_{\text{mean}} - \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} p_{ij}(a, \sigma^2) \right|.$$

In the end, the resulting probabilities have two values, one if $w_i = w_j$, and another one if $w_i \neq w_j$. Thus, the resulting models are two-community stochastic block models with two equally sized communities. We set $\lambda \in \{0, 0.5, 1, 1.5, ..., 4\}$. Note that $\lambda = 0$ represents the null model where all edge probabilities are $p_{ij} = p_{\text{mean}}$, $i < j = 1, ..., n$. Increasing values of λ yield models with rising heterogeneity: The intra-community connection probabilities increase, whereas the probabilities for inter-community edges decrease.

Performance results for $p_{\text{mean}} = \frac{1}{\sqrt{n}}$ are illustrated in Figure 17. Note that these results stand representative for all other values of p_{mean}, since the power of the tests is not directly influenced by the density of the underlying network. Overall, the results are in accordance with our theoretical analysis of Section 4.3.4. It is clearly visible that the tests based on C_3 have a superior performance in the underlying SBM cases as it is able to react way more sensitively than both other tests that are mainly based on P_3. Whereas C_3 is able to detect differences even for low values of the heterogeneity parameter λ, the two other tests require a rather large deviation from the null model. In this context, the performances get more reliable for all tests with an increasing amount of nodes. For the case of $n = 16$, P_3 and V_n are even not at all sensitive for all investigated values of λ. On a further note, the similar results of these two tests in all situations underline our derivations that P_3 is the decisive component of the centered subgraph count representation of V_n. The performance of the eigenvector centrality test is similar to these of $S_n(P_3)$ and V_n.

For our second SBM setup, we construct an SBM with more flexible assumptions to its parameters. Compared to before, it consists of three blocks with random block sizes and varying intra-block probabilities. Formally, we set this up by assigning the weight vector $\boldsymbol{w'}$ to

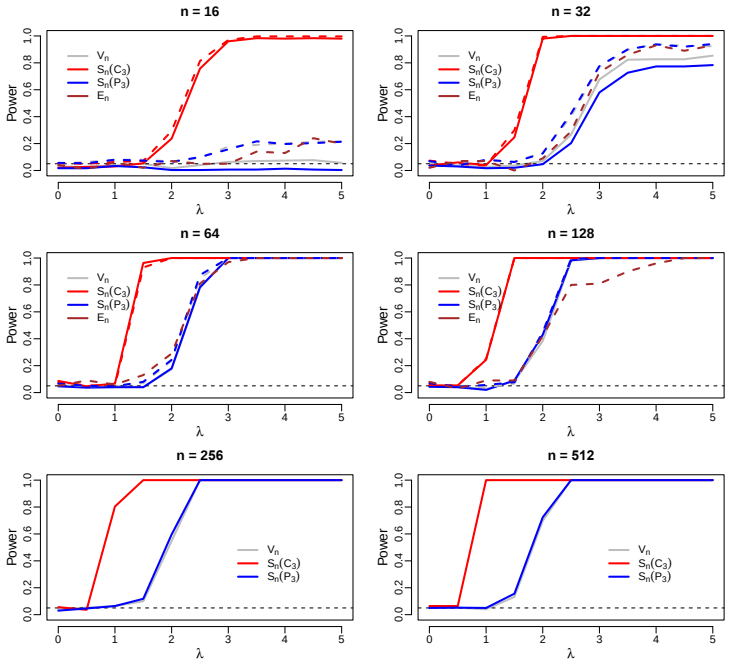


Figure 17: Power of the tests V_n, E_n, $S_n(C_3)$ and $S_n(P_3)$ for the alternative of an SBM with two blocks and $p_{\text{mean}} = \frac{1}{\sqrt{n}}$ for various values for n and λ. The asymptotic versions of the tests are represented by solid lines and the bootstrap version by dashed lines

the nodes with

$$\boldsymbol{w'} = (w'_1, ..., w'_n) = \Big(\underbrace{-\frac{1}{2}, ..., -\frac{1}{2}}_{b_1 \text{ times}}, \underbrace{0, ..., 0}_{b_2 \text{ times}}, \underbrace{\frac{1}{4}, ..., \frac{1}{4}}_{b_3 \text{ times}} \Big).$$

In this context, b_k with $k = \{1, 2, 3\}$ denotes the block size of the corresponding block. Each b_k is randomly chosen with the constraint that $b_1 + b_2 + b_3 = n$ and $b_k \geq 2$ for each k. The probabilities p_{ij} are then determined analogously to the two-block SBM case explained above.

The resulting performances are depicted in Figure 18 for $p_{\text{mean}} = \frac{\log n}{n}$. Regarding the performance under the null, all tests stick to the desired level $\alpha = 0.05$. Apart from this, the results are, interestingly, quite different to before as both tests based on P_3 perform superior compared to the test C_3 and also to E_n. Apparently, the two-star structure is more sensitive to an increased number of blocks, varying block sizes and varying intra-block probabilities. The bootstrap versions of the tests confirm this behaviour by achieving very similar results to their asymptotic counterparts. The improved performance of P_3 and V_n seems quite plausible. As shown in Section 4.3.4, subgraph counts of P_3 are, in theory, sensitive to detecting SBMs with the assumptions of Theorem 3. Although they are less sensitive compared to C_3 for the two block case, they could benefit from splitting the graph into more blocks as this can be interpreted as a further increase in heterogeneity compared to the ER model. Furthermore, the varying block sizes and intra-block probabilities might influence this behaviour as well.

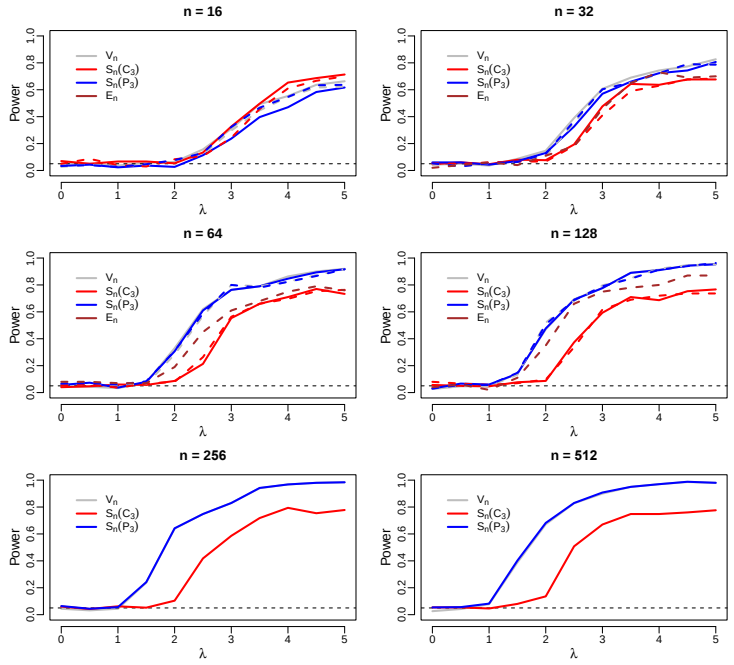


Figure 18: Power of the tests V_n, $S_n(C_3)$, E_n, and $S_n(P_3)$ for the alternative of an SBM with three blocks and $p_{\text{mean}} = \frac{\log n}{n}$. The asymptotic versions of the tests are represented by solid lines and the bootstrap version by dashed lines.

Interestingly, the test based on C_3 performs superior again for denser setups which can be seen in Figure 20 in Appendix B. In these denser networks, an increased λ enables the possibility of larger deviations between the intra-group probabilities and inter-group probabilities which could support the sensitivity of C_3.

Covariate models

As a further heterogenous alternative, we study covariate models for which the connection probabilities are generated in a way such that vertices with similar properties are more likely to connect than others. This characteristic seems to be a realistic setting for modeling real world networks. Especially in social networks, we expect variables such as the age or the social status to affect whether people know each other or not. To achieve a setting like this, we associate each vertex $v_i \in \{v_1, ..., v_n\}$ with a bivariate covariate

$$\begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix} \sim \mathcal{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right),$$

drawn independently from a bivariate standard normal distribution. The covariates are then multiplied by a sequence of variances $\sigma^2 \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2\}$. This yields versions of the same dataset with different degrees of heterogeneity. The connection probabilities

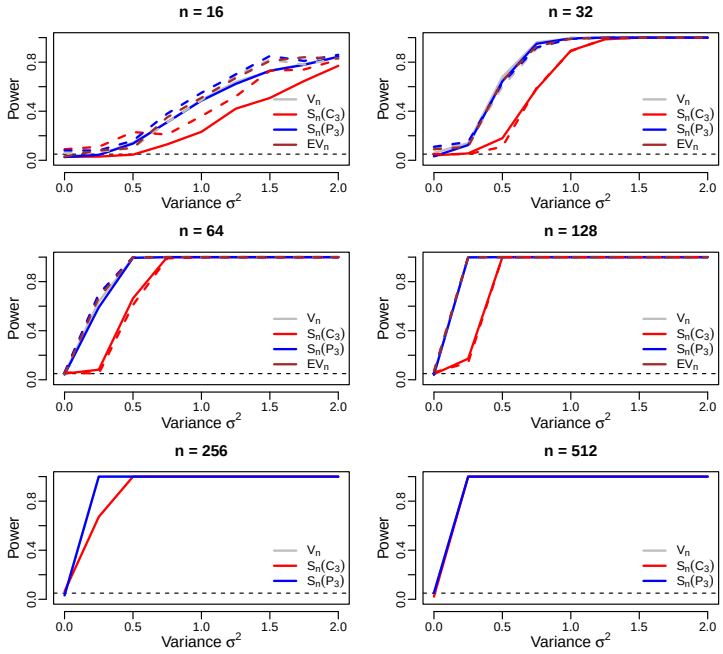


Figure 19: Power of the tests V_n, E_n, $S_n(C_3)$ and $S_n(P_3)$ for the alternative of a covariate model setup with $p_{\text{mean}} = \frac{1}{\sqrt{n}}$. The asymptotic versions of the tests are represented by solid lines and the bootstrap version by dashed lines.

(for given σ^2 and covariate set) are then calculated according to a logistic model:

$$p_{ij}(a, \sigma^2) = \begin{cases} \frac{\exp(a - \sigma^2|x_{i1} - x_{j1}| - \sigma^2|x_{i2} - x_{j2}|)}{1 + \exp(a - \sigma^2|x_{i1} - x_{j1}| - \sigma^2|x_{i2} - x_{j2}|)}, & \text{for } i \neq j, \\ 0, & \text{else.} \end{cases}$$

The constant a is again set to preserve the mean connectivity p_{mean} and can be determined as before. In this model, $\sigma^2 = 0$ represents the null hypothesis, each vertex has the same covariate value and the resulting connection probabilities are all equal to p_{mean}. The bigger the variance σ^2, the more the model deviates from the null model due to the increasing heterogeneity of the covariates. In general, the vertices have a higher probability to connect if their covariates have similar values. Especially, vertices that have unusually large or small covariate values will only be able to draw few connections. The described procedure yields a series of connection probability matrices with increasing heterogeneity.
Results are illustrated in Figure 19 for $p_{\text{mean}} = \frac{1}{\sqrt{n}}$. First of all, the tests reliably stick to the desired level $\alpha = 0.05$. Regarding the performance under the alternative, the performance is overall quite good for all tests with advantages for E_n, V_n and P_3 compared to C_3. For larger network sizes, all tests have no problems with detecting the deviation from the null model - even for the smallest investigated value of the heterogeneity parameter σ^2. Another result is that the bootstrap versions are able to outperform the asymptotic versions for small n and perform equally well for larger n. This behavior is also confirmed by the results of most of

the other investigated parameter setups that are not reported in the paper. As expected, the derived parametric bootstrap procedure is a promising alternative to the asymptotic version of the proposed class of tests – particularly for small network sizes.

4.6 Conclusion

Network data is quite complex and its handling is challenging from a statistical point of view. An important aspect is the complexity-reduction achieved by using a suitable network model. For a reliable analysis, however, the fitted model should be an adequate representation of the underlying data. In this context, we derived a class of goodness-of-fit tests for networks that serves as a unified approach and contains various formerly proposed tests as special cases. To do so, we used the broad class of graph functionals as test statistics and leveraged existing theory in order to derive novel asymptotic tests for evaluating if an underlying graph is generated by a homogenous Erdös-Rényi model or some heterogenous alternative. Moreover, we proposed a parametric bootstrap approach that particularly performs favorable in small network size situations compared to the asymptotic tests. We enriched our analysis by the application of our general procedure to three types of test statistics and derived power analysis results for the subgraphs of triangles and two-stars for the popular use case of stochastic block models. We illustrate our findings with an extensive simulation study in which we investigated multiple network parameter setups and also studied covariate models as a further heterogeneous alternative.

A possible topic for future research is to construct tests for which a heterogeneous model $\mathcal{G}_{\mathrm{HER}}(n)$ represents the null model in the spirit of Ouadah et al., 2020, who exclusively considered the degree variance and without any resampling-based inference. A possible approach could be to try to extend the proof technique for the $\mathcal{G}_{\mathrm{ER}}(n)$ case [Janson et al., 2011], that is based on a continuos time martingale theorem, to the heterogenous case yielding a heterogenous equivalent to the method of higher projections.

Appendix A: Proofs

Proof of Proposition 3

We have

$$\frac{X_n - \widehat{\mathbb{E}_{H_0}}(X_n)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n)}} = \frac{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(X_n)}}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n)}} \left(\frac{X_n - \mathbb{E}_{H_0}(X_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(X_n)}} + \frac{\mathbb{E}_{H_0}(X_n) - \widehat{\mathbb{E}_{H_0}}(X_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(X_n)}} \right).$$

Then, for the first factor on the last right-hand side, we have $\mathbb{V}\mathrm{ar}_{H_0}(X_n)/\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(X_n) \xrightarrow{P} 1$ due to condition (A) in (4.9). Further, according to condition (B) in (4.9), the second term in parentheses, that is, $(\mathbb{E}_{H_0}(X_n) - \widehat{\mathbb{E}_{H_0}}(X_n))/(\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(X_n)})$ vanishes as $n \to \infty$. Altogether, as the first term in parentheses, i.e. $(X_n - \mathbb{E}_{H_0}(X_n))/\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(X_n)}$ converges in distribution to $\mathcal{N}(0,1)$ according to Theorem 1, by Slutsky's Lemma, the claimed result follows. □

Proof of Proposition 4

To derive the rate of $\widehat{p} - p$, from $\mathbb{E}(\widehat{p}) = p$ and

$$\mathbb{V}\mathrm{ar}(\widehat{p}) = \binom{n}{2}^{-2} \sum_{1 \le i < j \le n} \mathbb{V}\mathrm{ar}(A_{ij}) = \binom{n}{2}^{-2} \sum_{1 \le i < j \le n} p(1-p),$$

under the null of a homogeneous ER graph, we get

$$\widehat{p} - p = O_P\left(\frac{\sqrt{p}}{n}\right).$$

Consequently, we also have

$$\begin{aligned}
\widehat{p}(1-\widehat{p}) - p(1-p) =& \widehat{p}(1-\widehat{p}) - \widehat{p}(1-p) + \widehat{p}(1-p) - p(1-p) = \widehat{p}(p-\widehat{p}) + (\widehat{p}-p)(1-p) \\
=& (\widehat{p} - p + p)(p-\widehat{p}) + (\widehat{p}-p)(1-p) = p(p-\widehat{p}) - (\widehat{p}-p)^2 + (\widehat{p}-p)(1-p) \\
=& \mathcal{O}_{\mathbb{P}}\left(\frac{p^{3/2}}{n}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{p}{n^2}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{\sqrt{p}}{n}\right) = \mathcal{O}_{\mathbb{P}}\left(\frac{\sqrt{p}}{n}\right)
\end{aligned}$$

We then consider the following expansion:

$$\frac{V_n - \widehat{\mathbb{E}_{H_0}}(V_n)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(V_n)}} = \frac{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(V_n)}}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(V_n)}} \left(\frac{V_n - \mathbb{E}_{H_0}(V_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(V_n)}} - \frac{\widehat{\mathbb{E}_{H_0}}(V_n) - \mathbb{E}_{H_0}(V_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(V_n)}} \right).$$

For the difference of the expectations in the numerator of the second term in brackets, we get

$$\widehat{\mathbb{E}_{H_0}}(V_n) - \mathbb{E}_{H_0}(V_n) = \frac{(n-1)(n-2)}{n}\Big(\widehat{p}(1-\widehat{p}) - p(1-p)\Big) = \mathcal{O}(n) \cdot \mathcal{O}_{\mathbb{P}}\left(\frac{\sqrt{p}}{n}\right) = \mathcal{O}_{\mathbb{P}}(\sqrt{p}).$$

As the variance $\mathbb{V}\mathrm{ar}_{H_0}(V_n)$ has order np^2, and $np \to \infty$ by assumption, we obtain

$$\frac{\widehat{\mathbb{E}_{H_0}}(V_n) - \mathbb{E}_{H_0}(V_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(V_n)}} = \mathcal{O}_{\mathbb{P}}\left(\frac{\sqrt{p}}{\sqrt{np^2}}\right) = \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{np}}\right) \xrightarrow{P} 0, \quad \text{as well as} \quad \frac{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(V_n)}}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(V_n)}} \xrightarrow{P} 1,$$

by similar arguments using $\widehat{p} - p = O_P(\frac{\sqrt{p}}{n})$. Combining these two convergence results and applying Slutzky's lemma, we get that $\frac{V_n - \widehat{\mathbb{E}_{H_0}}(V_n)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}_{H_0}}(V_n)}}$ has the same asymptotic distribution as $\frac{V_n - \mathbb{E}_{H_0}(V_n)}{\sqrt{\mathbb{V}\mathrm{ar}_{H_0}(V_n)}}$ stated in (4.13). □

Proof of Proposition 5

We have

$$\begin{aligned}
&\frac{\widehat{S}_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(\widehat{p}(1-\widehat{p}))^{m(H)}}} \\
=&\sqrt{\frac{\mathrm{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}{\mathrm{aut}(H)(n)_{n(H)}(\widehat{p}(1-\widehat{p}))^{m(H)}}}\left(\frac{S_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}} + \frac{\widehat{S}_n(H) - S_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}}\right) \\
=&\sqrt{\frac{(p(1-p))^{m(H)}}{(\widehat{p}(1-\widehat{p}))^{m(H)}}}\left(\frac{S_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}} + \frac{\widehat{S}_n(H) - S_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}}\right).
\end{aligned}$$

Hence, to prove the claimed result, by using the continous mapping theorem and by Slutsky's Lemma, it remains to show

$$\text{(a)} \quad \frac{\widehat{S}_n(H) - S_n(H)}{\sqrt{\mathrm{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}} \xrightarrow{P} 0, \quad \text{and} \quad \text{(b)} \quad \frac{(p(1-p))^{m(H)}}{(\widehat{p}(1-\widehat{p}))^{m(H)}} \xrightarrow{P} 1.$$

For showing part (b), it is equivalent to show that the reciprocal, that is,

$$\frac{(\widehat{p}(1-\widehat{p}))^{m(H)}}{(p(1-p))^{m(H)}} = \left(\frac{\widehat{p}}{p}\right)^{m(H)}\left(\frac{1-\widehat{p}}{1-p}\right)^{m(H)},$$

converges to 1. As $\widehat{p} - p = O_P(\frac{\sqrt{p}}{n})$, we have $\widehat{p}/p - 1 = O_P(\frac{1}{\sqrt{p}n}) = o_P(1)$ as $n \to \infty$ due to $np^{2/3} \to \infty$ by assumption. Consequently, by continuous mapping, we have also $(p/\widehat{p})^{m(H)} - 1^{m(H)} = (p/\widehat{p})^{m(H)} - 1 = o_P(1)$. Similarly, we have

$$\frac{1-\widehat{p}}{1-p} = \frac{1-p}{1-p} + \frac{p-\widehat{p}}{1-p} = 1 + \frac{p-\widehat{p}}{1-p} = 1 + O_P\left(\frac{\sqrt{p}}{n}\right) = 1 + o_P(1),$$

because $p = p(n) \to p_0 \in [0,1)$. Hence, by continuous mapping, we have $((1-p)/(1-\widehat{p}))^{m(H)} - 1^{m(H)} = ((1-p)/(1-\widehat{p}))^{m(H)} - 1 = o_P(1)$ as well. Altogether, we proved part (b). For part

(a), under the null of an ER graph, it is natural to decompose $\widehat{S}_n(H)$ and to re-write it in terms of centered subgraph counts. For this purpose, let G be a graph on n vertices, and H be another graph with $n(H) \leq n$. Then, similar to Example 6.50 in Janson et al., 2011, let $J \subseteq H$ be a graph that is obtained from H by dropping some of the edges present in H. Hence, we have $n(J) = n(H)$, but $m(J) \leq m(H)$. We consider the $(n)_{n(J)}$ different injective mappings $\varphi_{n(J)}$ from vertices of J into $\{1, ..., n\}$. Each mapping $\varphi_{n(J)}$ maps J onto a copy $\varphi_{n(j)}(J)$ of J in a complete graph with n vertices. Then, for the feasible centered subgraph count $\widehat{S}_n(H)$, we get

$$\begin{aligned}
\widehat{S}_n(H) &= \sum_{\varphi} \prod_{e \in E(\varphi(H))} (\mathbb{1}\{e \in E(G)\} - \widehat{p}) = \sum_{\varphi_{n(H)}} \prod_{e \in E(\varphi_{n(H)}(H))} (\mathbb{1}\{e \in E(G)\} - p + p - \widehat{p}) \qquad (4.21) \\
&= \sum_{\substack{J \subseteq H \\ J \text{ induced}}} \left[\sum_{\varphi_{n(J)}} \prod_{e \in E_{n(J)}(\varphi_{n(J)}(H))} (\mathbb{1}\{e \in E(G)\} - p) \right] (p - \widehat{p})^{m(H)-m(J)} \\
&= \sum_{\substack{J \subseteq H \\ J \text{ induced}}} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) \, (p - \widehat{p})^{m(H)-m(J)},
\end{aligned}$$

where $k(J)$ denotes the number of isolated vertices in J, $\widetilde{J}$ denotes the graph obtained from J after removing all isolated vertices, $\prod_{h=n(H)}^{n(H)-1}(n-h) := 1$ and the sum $\sum_{J \subseteq H, J \text{ induced}}$ ranges over all substructures that can be obtained from H by dropping j, $j = 0, 1, \ldots, m(H)$ edges of the edges present in H. Separating the case, where no edge is dropped at all, we get

$$\widehat{S}_n(H) = S_n(H) + \sum_{\substack{J \subsetneq H \\ J \text{ induced}}} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) \, (p - \widehat{p})^{m(H)-m(J)},$$

where $m(H) - m(J) \geq 1$ by construction. For the sake of readability, we suppress the expression "J is induced" in the index of the sum in the remainder of this proof. As we assumed $H \neq \emptyset$, $H \neq P_2$, using Proposition 1 as well as $S_n(\emptyset) := 1$, we get

$$\begin{aligned}
&\frac{\widehat{S}_n(H) - S_n(H)}{\sqrt{\text{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}} = \frac{\sum_{J \subsetneq H} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) \, (p - \widehat{p})^{m(H)-m(J)}}{\sqrt{\text{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}} \\
=&\frac{\sum_{\substack{J \subsetneq H \\ m(J)>0}} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) \, (p - \widehat{p})^{m(H)-m(J)}}{\sqrt{\text{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}} \\
+&\frac{\sum_{\substack{J \subsetneq H \\ m(J)=0}} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) \, (p - \widehat{p})^{m(H)-m(J)}}{\sqrt{\text{aut}(H)(n)_{n(H)}(p(1-p))^{m(H)}}} \\
=&I + II
\end{aligned}$$

with an obvious notation for I and II. For the term I, using Proposition 1, which holds for non-empty graphs, we get

$$I = O_P\left(\frac{\sum_{\substack{J\subsetneq H\\ m(J)>0}} n^{k(J)} n^{n(\widetilde{J})/2} p^{m(\widetilde{J})/2} \left(\frac{\sqrt{p}}{n}\right)^{m(H)-m(J)}}{\sqrt{n^{n(H)}(p(1-p))^{m(H)}}}\right)$$

Further, as $m(\widetilde{J}) = m(J)$ and $n(\widetilde{J}) = n(H) - k(J)$ holds by construction, the last right-hand side becomes

$$O_P\left(\frac{\sum_{\substack{J\subsetneq H\\ m(J)>0}} n^{k(J)} n^{\frac{n(H)-k(J)}{2}} p^{m(J)/2} \left(\frac{\sqrt{p}}{n}\right)^{m(H)-m(J)}}{\sqrt{n^{n(H)} p^{m(H)}}}\right) = O_P\left(\frac{\sum_{\substack{J\subsetneq H\\ m(J)>0}} n^{\frac{n(H)+k(J)}{2}} p^{\frac{m(J)}{2}} p^{\frac{m(H)-m(J)}{2}} n^{-m(H)+m(J)}}{\sqrt{n^{n(H)} p^{m(H)}}}\right)$$

$$= O_P\left(\frac{\sum_{\substack{J\subsetneq H\\ m(J)>0}} n^{\frac{n(H)}{2}} n^{\frac{k(J)}{2}} p^{\frac{m(H)}{2}} n^{-m(H)+m(J)}}{\sqrt{n^{n(H)} p^{m(H)}}}\right) = O_P\left(\sum_{\substack{J\subsetneq H\\ m(J)>0}} n^{-m(H)+m(J)} n^{\frac{k(J)}{2}}\right).$$

Now, as $m(J) > 0$ and H is connected, we have $k(J) \leq m(H) - m(J)$. Hence, as for all J with $J \subsetneq H$, we have $m(H) - m(J) \geq 1$ and as the sum $\sum_{J\subsetneq H}$ is finite, because $n(H)$ is finite, the last right-hand side is bounded by

$$O_P\left(\sum_{\substack{J\subsetneq H\\ m(J)>0}} n^{-m(H)+m(J)} \sqrt{n^{m(H)-m(j)}}\right) = O_P\left(\sum_{\substack{J\subsetneq H\\ m(J)>0}} \sqrt{\left(\frac{1}{n}\right)^{m(H)-m(J)}}\right) \leq O_P\left(\frac{1}{\sqrt{n}}\right) = o_P(1).$$

For the term II, due to $m(J) = 0$ implying $J = \emptyset$ and using $S_n(\emptyset) := 1$, we get

$$II = O_P\left(\frac{n^{n(H)} \left(\frac{\sqrt{p}}{n}\right)^{m(H)}}{\sqrt{n^{n(H)}(p(1-p))^{m(H)}}}\right) = O_P\left(n^{n(H)/2-m(H)}\right) = o_P(1)$$

as $H \neq P_2$ and H connected leading to $n(H) \geq 3$ and $m(H) \geq n(H) - 1$. □

Proof of Proposition 6

Following the steps in (4.21), we get

$$\begin{aligned}
T_n(H) =& \frac{1}{\mathrm{aut}(H)} \sum_{\varphi} \prod_{e \in E(\varphi(H))} (\mathbb{1}\{e \in E(G)\}) = \frac{1}{\mathrm{aut}(H)} \sum_{\varphi_{n(H)}} \prod_{e \in E(\varphi_{n(H)}(H))} (\mathbb{1}\{e \in E(G)\} - p + p) \\
=& \frac{1}{\mathrm{aut}(H)} \sum_{\substack{J \subseteq H \\ J \text{ induced}}} \left[\sum_{\varphi_{n(J)}} \prod_{e \in E_{n(J)}(\varphi(H))} (\mathbb{1}\{e \in E(G)\} - p) \right] p^{m(H)-m(J)} \\
=& \frac{1}{\mathrm{aut}(H)} \sum_{\substack{J \subseteq H \\ J \text{ induced}}} \left(\prod_{h=n(H)-k(J)}^{n(H)-1} (n-h) \right) S_n(\widetilde{J}) p^{m(H)-m(J)}.
\end{aligned}$$

□

Proof of Theorem 2

Starting with the expected value of $S_n(C_3)$, as all edges form independently, we get

$$\begin{aligned}
\mathbb{E}(S_n(C_3)) =& \sum_{\varphi} \mathbb{E}\Big(\prod_{e \in E(\varphi(C_3))} \Big(\mathbb{1}\{e \in E(G)\} - p \Big) \Big) = \sum_{\varphi} \prod_{e \in E(\varphi(C_3))} \mathbb{E}\Big(\Big(\mathbb{1}\{e \in E(G)\} - p \Big) \Big) \\
=& \sum_{\varphi} \prod_{e \in E(\varphi(C_3))} \Big(\mathbb{P}\{e \in E(G)\} - p \Big).
\end{aligned}$$

Obviously, for the $\mathcal{G}_{\mathrm{ER}}(n)$ model, we get $\mathbb{E}_{\mathrm{ER}}(S_n(C_3)) = \sum_{\varphi} \prod_{e \in E(\varphi(C_3))} \Big(p - p \Big) = 0$ by construction, where $\mathbb{P}\{e \in E(G)\} = p$. In contrast, for the SBM case, we have

$$\begin{aligned}
\mathbb{E}_{\mathrm{SBM}}(S_n(C_3)) =& \sum_{\varphi} \prod_{e \in E(\varphi(C_3))} \Big(\mathbb{P}\{e \in E(G)\} - p \Big) \\
&= n\Big(\frac{n}{2} - 1\Big)\Big(\frac{n}{2} - 2\Big)(p_{\mathrm{intra}} - p)^3 + n\Big(\frac{n}{2} - 1\Big)\Big(\frac{n}{2}\Big)(p_{\mathrm{intra}} - p)(p_{\mathrm{inter}} - p)^2 \\
&+ n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2} - 1\Big)(p_{\mathrm{inter}} - p)(p_{\mathrm{intra}} - p)(p_{\mathrm{inter}} - p) + n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2} - 1\Big)(p_{\mathrm{inter}} - p)^2(p_{\mathrm{intra}} - p) \\
&= n\Big(\frac{n}{2} - 1\Big)\Big(\frac{n}{2} - 2\Big)(p_{\mathrm{intra}} - p)^3 + 3n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2} - 1\Big)(p_{\mathrm{intra}} - p)(p_{\mathrm{inter}} - p)^2.
\end{aligned}$$

Next, we express $p = p_{mean}$ in terms of p_{intra} and p_{inter}, which results in $p = \frac{n-2}{2n-2}p_{\text{intra}} + \frac{n}{2n-2}p_{\text{inter}}$. Therefore, the last right-hand side becomes

$$\begin{aligned}
&n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\Big(p_{\text{intra}} - \Big(\frac{n-2}{2n-2}p_{\text{intra}} + \frac{n}{2n-2}p_{\text{inter}}\Big)\Big)^3 \\
&+ 3n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\Big(p_{\text{intra}} - \Big(\frac{n-2}{2n-2}p_{\text{intra}} + \frac{n}{2n-2}p_{\text{inter}}\Big)\Big)\Big(p_{\text{inter}} - \Big(\frac{n-2}{2n-2}p_{\text{intra}} + \frac{n}{2n-2}p_{\text{inter}}\Big)\Big)^2 \\
=&n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\Big(\frac{n}{2n-2}p_{\text{intra}} - \frac{n}{2n-2}p_{\text{inter}}\Big)^3 \\
&+ 3n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2n-2}p_{\text{intra}} - \frac{n}{2n-2}p_{\text{inter}}\Big)\Big(\frac{n-2}{2n-2}p_{\text{inter}} - \frac{n-2}{2n-2}p_{\text{intra}}\Big)^2,
\end{aligned}$$

which is asymptotically equivalent to

$$\begin{aligned}
&n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\frac{1}{8}(p_{\text{intra}} - p_{\text{inter}})^3 + 3n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\frac{1}{8}(p_{\text{intra}} - p_{\text{inter}})(p_{\text{inter}} - p_{\text{intra}})^2 \\
&= n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\frac{1}{8}\epsilon^3 + 3n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\frac{1}{8}\epsilon(-\epsilon)^2 \\
&= \frac{1}{8}n\Big(\frac{n}{2}-1\Big)\left[\frac{n}{2}-2+\frac{3}{2}n\right]\epsilon^3 = \frac{n\left(\frac{n}{2}-1\right)(2n-2)}{8}\epsilon^3 = \frac{n(n-1)(n-2)}{8}\epsilon^3 = O(n^3\epsilon^3).
\end{aligned}$$

With the help of Proposition 1, we can determine the variance for the ER-model as

$$\mathbb{V}\text{ar}_{\text{ER}}(S_n(C_3)) = 6n(n-1)(n-2)p^3(1-p)^3 = 36\binom{n}{3}p^3(1-p)^3 = O(n^3p^3).$$

Altogether, asymptotically, we get

$$\frac{\mathbb{E}_{\text{SBM}}(S_n(C_3)) - \mathbb{E}_{\text{ER}}(S_n(C_3))}{\sqrt{\mathbb{V}\text{ar}_{\text{ER}}(S_n(C_3))}} = O\left(\frac{n^3\epsilon^3}{\sqrt{n^3p^3}}\right).$$

Considering the case of fixed alternatives with $\epsilon \sim p$, we get the following rates: if $p \to p_0 \in (0,1)$, we get an $O(n^{3/2})$. If $p \to 0$, we get an $O((np)^{3/2})$. Hence, if $np \to \infty$, we get divergence to $+\infty$ in all cases. For the case of local alternatives, where $\epsilon = \epsilon_n$ is allowed to decrease to 0 at a faster rate than $p = p_n$, we get the following: if $p \to p_0 \in (0,1)$, we get an $O(n^{3/2}\epsilon^3)$. If $p \to 0$, we get an $O(n^{3/2}\epsilon^3p^{-3/2})$.

Further, continuing with the expectation of $T_n(C_3)$, we have

$$\mathbb{E}(T_n(C_3)) = \frac{1}{\text{aut}(C_3)}\sum_{\varphi}\prod_{e\in E(\varphi(C_3))}\mathbb{P}\{e \in E(G)\}.$$

For the $\mathcal{G}_{\text{ER}}(n)$ model, this obviously simplifies to

$$\mathbb{E}_{\text{ER}}(T_n(C_3)) = \frac{1}{6}n(n-1)(n-2)p^3 = \binom{n}{3}p^3.$$

For the SBM case, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{SBM}}(T_n(C_3)) \\
=&\frac{1}{6}\left[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)p_{\mathrm{intra}}^3+3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)p_{\mathrm{intra}}p_{\mathrm{inter}}^2\right] \\
=&\frac{1}{6}\left[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)(p_{\mathrm{intra}}-p+p)^3+3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)(p_{\mathrm{intra}}-p+p)(p_{\mathrm{inter}}-p+p)^2\right] \\
=&\frac{1}{6}\Big\{n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\left[(p_{\mathrm{intra}}-p)^3+3(p_{\mathrm{intra}}-p)^2p+3(p_{\mathrm{intra}}-p)p^2+p^3\right] \\
&+3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\left((p_{\mathrm{intra}}-p)\left[(p_{\mathrm{inter}}-p)^2+2(p_{\mathrm{inter}}-p)p+p^2\right]+p\left[(p_{\mathrm{inter}}-p)^2+2(p_{\mathrm{inter}}-p)p+p^2\right]\right)\Big\}.
\end{aligned}
$$

Analogously to before, we express p in terms of p_{intra} and p_{inter}, which results in $p=\frac{n-2}{2n-2}p_{\mathrm{intra}}+\frac{n}{2n-2}p_{\mathrm{inter}}$. This is asymptotically equal to $p=\frac{1}{2}p_{\mathrm{intra}}+\frac{1}{2}p_{\mathrm{inter}}$. Hence, we have

$$
\begin{aligned}
&\frac{1}{6}\Bigg\{n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\left[\frac{1}{8}\epsilon^3+\frac{3}{4}\epsilon^2p+\frac{3}{2}\epsilon p^2+p^3\right] \\
&+3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\left(\frac{1}{2}\epsilon\left[\frac{1}{4}\epsilon^2+(-\epsilon)p+p^2\right]+p\left[\frac{1}{4}(\epsilon)^2+2(-\epsilon)p+p^2\right]\right)\Bigg\} \\
=&\frac{1}{6}\left\{n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\left[\frac{1}{8}\epsilon^3+\frac{3}{4}\epsilon^2p+\frac{3}{2}\epsilon p^2+p^3\right]+3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\left[\frac{1}{8}\epsilon^3-\frac{1}{2}\epsilon^2p+\frac{1}{2}\epsilon p^2+\frac{1}{4}\epsilon^2p-\epsilon p^2+p^3\right]\right\}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{6}\Big\{\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)+3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big]\frac{1}{8}\epsilon^3+\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\frac{3}{4}-\frac{3}{4}n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big]\epsilon^2p \\
&+\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\frac{3}{2}-\frac{3}{2}n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big]\epsilon p^2+\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)+3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big]p^3\Big\} \\
=&\frac{1}{6}\Big\{n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2+\frac{3}{2}n\Big)\frac{1}{8}\epsilon^3+n\Big(\frac{n}{2}-1\Big)\frac{3}{4}\Big(\frac{n}{2}-2-\frac{n}{2}\Big)\epsilon^2p+n\Big(\frac{n}{2}-1\Big)\frac{3}{2}\Big(\frac{n}{2}-2-\frac{n}{2}\Big)\epsilon p^2 \\
&+n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2+\frac{3}{2}n\Big)p^3\Big\} \\
=&\frac{1}{6}\left[\frac{n(n-1)(n-2)}{8}\epsilon^3-\frac{3n(n-2)}{4}\epsilon^2p-\frac{3n(n-2)}{2}\epsilon p^2+n(n-1)(n-2)p^3\right].
\end{aligned}
$$

Hence, it yields

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{SBM}}(T_n(C_3))-\mathbb{E}_{\mathrm{ER}}(T_n(C_3)) \\
&=\frac{1}{6}\left[\frac{n(n-1)(n-2)}{8}\epsilon^3-\frac{3n(n-2)}{4}\epsilon^2p-\frac{3n(n-2)}{2}\epsilon p^2+n(n-1)(n-2)p^3\right]-\binom{n}{3}p^3 \\
&=\frac{1}{6}\left[\frac{n(n-1)(n-2)}{8}\epsilon^3-\frac{3n(n-2)}{4}\epsilon^2p-\frac{3n(n-2)}{2}\epsilon p^2\right]=O(n^3\epsilon^3)+O(n^2\epsilon^2p)+O(n^2\epsilon p^2).
\end{aligned}
$$

With (4.15) from Proposition 6, we get

$$
\mathbb{V}\mathrm{ar}_{\mathrm{ER}}(T_n(C_3))=\binom{n}{3}p^3(1-p)^3+3\binom{n}{3}p^4(1-p)^2+3(n-2)\binom{n}{3}p^5(1-p).
$$

Altogether, asymptotically, we get

$$\frac{\mathbb{E}_{\mathrm{SBM}}(T_n(C_3)) - \mathbb{E}_{\mathrm{ER}}(T_n(C_3))}{\sqrt{\mathbb{V}\mathrm{ar}_{\mathrm{ER}}(T_n(C_3))}} = O\left(\frac{n^3\epsilon^3 + n^2\epsilon^2 p + n^2\epsilon p^2}{\sqrt{n^3p^3 + n^3p^4 + n^4p^5}}\right).$$

Considering the case of fixed alternatives with $\epsilon \sim p$, we get the following rates: if $p \to p_0 \in (0,1)$, we get an $O(n)$. If $p \to 0$ and $np^2 \to \infty$, we get an $O(n\sqrt{p})$. If $p \to 0$ and $np^2 \to K \geq 0$, we get an $O((np)^{3/2})$. Hence, if $np \to \infty$, we get divergence to $+\infty$ in all three cases. For the case of local alternatives, where $\epsilon = \epsilon_n$ is allowed to decrease to 0 at a faster rate than $p = p_n$, we get the following: if $n\epsilon \to \infty$ and $np^2 \to \infty$, we get an $O(n\epsilon^3 p^{-5/2})$, if $n\epsilon \to K_\epsilon \geq 0$ and $np^2 \to \infty$, we get an $O(\epsilon p^{-5/2})$. If $n\epsilon \to \infty$ and $np^2 \to K \geq 0$, we get an $O(n^{3/2}\epsilon^3 p^{-3/2})$, if $n\epsilon \to K_\epsilon \geq 0$ and $np^2 \to K \geq 0$, we get an $O(n^{1/2}\epsilon p^{-3/2})$.

Proof of Theorem 3

Analogously to the proof of Theorem 2, we have

$$\mathbb{E}(S_n(P_3)) = \sum_{\varphi} \prod_{e \in E(\varphi(P_3))} \Big(\mathbb{P}\{e \in E(G)\} - p\Big)$$

Again, it is $\mathbb{E}_{\mathrm{ER}}(S_n(P_3)) = \sum_\varphi \prod_{e\in E(\varphi(P_3))} \Big(p - p\Big) = 0$ for the $\mathcal{G}_{\mathrm{ER}}(n)$ model. For the SBM case, we have

$$\begin{aligned}
\mathbb{E}_{\mathrm{SBM}}(S_n(P_3)) =& n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)(p_{\mathrm{intra}}-p)^2 + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)(p_{\mathrm{inter}}-p)^2 \\
&+ n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)(p_{\mathrm{inter}}-p)(p_{\mathrm{intra}}-p) + n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)(p_{\mathrm{intra}}-p)(p_{\mathrm{inter}}-p) \\
=& n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)(p_{\mathrm{intra}}-p)^2 + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)(p_{\mathrm{inter}}-p)^2 \\
&+ 2n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)(p_{\mathrm{intra}}-p)(p_{\mathrm{inter}}-p)
\end{aligned}$$

We again express p of the ER model in terms of p_{intra} and p_{inter}, which results in $p = \frac{n-2}{2n-2}p_{\mathrm{intra}} + \frac{n}{2n-2}p_{\mathrm{inter}}$. The last right-hand side therefore is

$$\begin{aligned}
=& n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\Big(p_{\mathrm{intra}} - \Big(\frac{n-2}{2n-2}p_{\mathrm{intra}} + \frac{n}{2n-2}p_{\mathrm{inter}}\Big)\Big)^2 \\
&+ n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big(p_{\mathrm{inter}} - \Big(\frac{n-2}{2n-2}p_{\mathrm{intra}} + \frac{n}{2n-2}p_{\mathrm{inter}}\Big)\Big)^2 \\
&+ 2n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\Big(p_{\mathrm{intra}} - \Big(\frac{n-2}{2n-2}p_{\mathrm{intra}} + \frac{n}{2n-2}p_{\mathrm{inter}}\Big)\Big)\Big(p_{\mathrm{inter}} - \Big(\frac{n-2}{2n-2}p_{\mathrm{intra}} + \frac{n}{2n-2}p_{\mathrm{inter}}\Big)\Big) \\
=& n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\Big(\frac{n}{2n-2}p_{\mathrm{intra}} - \frac{n}{2n-2}p_{\mathrm{inter}}\Big)^2 + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big(\frac{n-2}{2n-2}p_{\mathrm{inter}} - \frac{n-2}{2n-2}p_{\mathrm{intra}}\Big)^2 \\
&+ 2n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2n-2}p_{\mathrm{intra}} - \frac{n}{2n-2}p_{\mathrm{inter}}\Big)\Big(\frac{n-2}{2n-2}p_{\mathrm{inter}} - \frac{n-2}{2n-2}p_{\mathrm{intra}}\Big),
\end{aligned}$$

which is asymptotically equivalent to

$$
\begin{aligned}
&= n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\frac{1}{4}(p_{\text{intra}}-p_{\text{inter}})^2 \\
&+ n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\frac{1}{4}(p_{\text{inter}}-p_{\text{intra}})^2 + 2n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\frac{1}{2}(p_{\text{intra}}-p_{\text{inter}})\frac{1}{2}(p_{\text{inter}}-p_{\text{intra}}) \\
&= n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\frac{1}{4}\epsilon^2 + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\frac{1}{4}(-\epsilon)^2 + 2n\Big(\frac{n}{2}\Big)\Big(\frac{n}{2}-1\Big)\frac{1}{2}\epsilon\frac{1}{2}(-\epsilon) \\
&= \frac{n(\frac{n}{2}-1)}{4}\left[\frac{n}{2}-2+\frac{n}{2}-2\frac{n}{2}\right]\epsilon^2 \\
&= \frac{n(\frac{n}{2}-1)}{4}(-2)\epsilon^2 = -\frac{n(n-2)}{4}\epsilon^2 = O(n^2\epsilon^2).
\end{aligned}
$$

With the help of Proposition 1, we can determine the variance for the ER-model as

$$\mathbb{V}\text{ar}_{\text{ER}}(S_n(P_3)) = 2n(n-1)(n-2)p^2(1-p)^2 = 12\binom{n}{3}p^2(1-p)^2 = O(n^3p^2)$$

Altogether, asymptotically, we get

$$\frac{\mathbb{E}_{\text{SBM}}(S_n(P_3)) - \mathbb{E}_{\text{ER}}(S_n(P_3))}{\sqrt{\mathbb{V}\text{ar}_{\text{ER}}(S_n(P_3))}} = O\left(\frac{n^2\epsilon^2}{\sqrt{n^3p^2}}\right)$$

Considering the case of fixed alternatives with $\epsilon \sim p$, we get the following rates: if $p \to p_0 \in (0,1)$, we get an $O(n^{1/2})$. If $p \to 0$, we get an $O(n^{1/2}p)$. Hence, if $np^2 \to \infty$, we get divergence to $+\infty$ in all cases. For the case of local alternatives, where $\epsilon = \epsilon_n$ is allowed to decrease to 0 at a faster rate than $p = p_n$, we get the following: if $p \to p_0 \in (0,1)$, we get an $O(n^{1/2}\epsilon^2)$. If $p \to 0$, we get an $O(n^{1/2}\epsilon^2 p^{-1})$.

Further, let us now move on to the raw subgraph counts. It yields

$$\mathbb{E}(T_n(P_3)) = \frac{1}{\text{aut}(P_3)}\sum_{\varphi}\prod_{e\in E(\varphi(C_3))}\mathbb{P}\{e \in E(G)\}.$$

For the $\mathcal{G}_{\text{ER}}(n)$ model, this obviously simplifies to

$$\mathbb{E}_{\text{ER}}(T_n(P_3)) = \frac{1}{2}n(n-1)(n-2)p^2 = 3\binom{n}{3}p^2.$$

For the SBM case, we have

$$
\begin{aligned}
&\mathbb{E}_{\mathrm{SBM}}(T_n(P_3)) \\
=&\frac{1}{2}\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)p_{\mathrm{intra}}^2 + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)p_{\mathrm{inter}}^2 + 2n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)p_{\mathrm{intra}}p_{\mathrm{inter}}\Big] \\
=&\frac{1}{2}\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)(p_{\mathrm{intra}}-p+p)^2 + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)(p_{\mathrm{inter}}-p+p)^2 \\
&+2n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)(p_{\mathrm{intra}}-p+p)(p_{\mathrm{inter}}-p+p)\Big] \\
=&\frac{1}{2}\Big\{n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\left[(p_{\mathrm{intra}}-p)^2+2(p_{\mathrm{intra}}-p)p+p^2\right] + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\left[(p_{\mathrm{inter}}-p)^2+2(p_{\mathrm{inter}}-p)p+p^2\right] \\
&+2n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\left[(p_{\mathrm{inter}}-p)(p_{\mathrm{intra}}-p)+(p_{\mathrm{intra}}-p)p+(p_{\mathrm{inter}}-p)p+p^2\right]\Big\}.
\end{aligned}
$$

Again, we express some p in terms of p_{intra} and p_{inter}, which results in $p = \frac{n-2}{2n-2}p_{\mathrm{intra}} + \frac{n}{2n-2}p_{\mathrm{inter}}$. This is asymptotically equivalent to $p = \frac{1}{2}p_{\mathrm{intra}} + \frac{1}{2}p_{\mathrm{inter}}$. Hence, we have

$$
\begin{aligned}
&\frac{1}{2}\Big\{n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big)\left[\frac{1}{4}\epsilon^2+\epsilon p+p^2\right] + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\left[-\frac{1}{4}\epsilon^2-\epsilon p+p^2\right] + 2n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\left[-\frac{1}{4}\epsilon^2+p^2\right]\Big\} \\
&= \frac{1}{2}\Big\{\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big) - n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big]\frac{1}{4}\epsilon^2 \\
&+\Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big) - n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big]\epsilon p + \Big[n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2\Big) + 3n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}\Big)\Big]p^2\Big\} \\
&= \frac{1}{2}\Big\{n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2+\frac{n}{2}-\frac{2n}{2}\Big)\frac{1}{4}\epsilon^2 + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2-\frac{n}{2}\Big)\epsilon p + n\Big(\frac{n}{2}-1\Big)\Big(\frac{n}{2}-2+\frac{3}{2}n\Big)p^2\Big\} \\
&= \frac{1}{2}\left[-\frac{n(n-2)}{4}\epsilon^2 - n(n-2)\epsilon p + n(n-1)(n-2)p^2\right].
\end{aligned}
$$

Hence, it yields

$$
\begin{aligned}
\mathbb{E}_{\mathrm{SBM}}(T_n(P_3)) - \mathbb{E}_{\mathrm{ER}}(T_n(P_3)) &= \frac{1}{2}\left[-\frac{n(n-2)}{4}\epsilon^2 - n(n-2)\epsilon p + n(n-1)(n-2)p^2\right] - 3\binom{n}{3}p^2 \\
&= \frac{1}{2}\left[-\frac{n(n-2)}{4}\epsilon^2 - n(n-2)\epsilon p\right] \\
&= O(n^2\epsilon^2) + O(n^2\epsilon p).
\end{aligned}
$$

Together with (4.15) from Proposition 6, we get

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}_{\mathrm{ER}}(T_n(P_3)) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{2}\left[S_n(P_3) + (n-2)S_n(P_2)p + n(n-1)(n-2)S_n(\emptyset)p^2\right]\right) \\
&= \frac{1}{2}n(n-1)(n-2)p^2(1-p)^2 + \frac{1}{2}(n-2)^2n(n-1)p^3(1-p) \\
&= 3\binom{n}{3}p^2(1-p)^2 + 3(n-2)\binom{n}{3}p^3(1-p) = O(n^3p^2) + O(n^4p^3).
\end{aligned}
$$

Altogether, asymptotically, we get

$$\frac{\mathbb{E}_{\mathrm{SBM}}(T_n(P_3)) - \mathbb{E}_{\mathrm{ER}}(T_n(P_3))}{\sqrt{\mathbb{V}\mathrm{ar}_{\mathrm{ER}}(T_n(P_3))}} = O\left(\frac{n^2\epsilon^2 + n^2\epsilon p}{\sqrt{n^3p^2 + n^4p^3}}\right)$$

Considering the case of fixed alternatives with $\epsilon \sim p$, we get the following rates: if $p \to p_0 \in (0,1)$, we get an $O(1)$. If $p \to 0$ and $np \to \infty$, we get an $O(\sqrt{p})$. If $p \to 0$ and $np \to K \geq 0$, we get an $O(\sqrt{n}p)$. Hence, we do not get divergence to $+\infty$ in all cases. For the case of local alternatives, where $\epsilon = \epsilon_n$ is allowed to decrease to 0 at a faster rate than $p = p_n$, we get the following: if $p \to p_0 \in (0,1)$, we get $O(\epsilon)$, if $p \to 0$ and $np \to \infty$, we get $O(\epsilon p^{-1/2})$, and if $p \to 0$ and $np \to K \geq 0$, we get $O(n^{1/2}\epsilon)$.

Proof of Theorem 4

By assumption, $G \in \mathcal{G}_{HER}$, that is, G is generated by a heterogeneous ER-graph $\mathcal{G}(n, \boldsymbol{P})$ with some $\boldsymbol{P} = (p_{ij})$, $p_{ij} = p_{ji}$, $p_{ij} \in [0,1)$ for all i, j such that its mean connectivity $p_{\mathrm{mean}} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} p_{n,ij}$ satisfies $p_{\mathrm{mean}} = p_{\mathrm{mean}}(n) \xrightarrow[n\to\infty]{} p_0 \in [0,1]$. Then, conditional on $\boldsymbol{A} = (A_{ij})_{1 \leq i,j \leq n}$, G^* is generated by a homogeneous ER-graph $\mathcal{G}(n, \widehat{p})$, where $\widehat{p} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} A_{ij}$. Furthermore, we have that, under the null, $X_n = \sum_{H \in \mathcal{H}} a_n(H) S_n(H)$, which is dominated by some family of connected graphs $\mathcal{H}_0$ such that for all $H \in \mathcal{H}_0$, it holds $np_{\mathrm{mean}}^{r(H)} \xrightarrow[n\to\infty]{} \infty$. Hence, under both the null and the alternative, we immediately get $X_n^* = \sum_{H \in \mathcal{H}_0} a_n^*(H) S_n^*(H)$, which is dominated by the same family of connected graphs $\mathcal{H}_0$, and where $a_n^*(H) = a_n(H, \widehat{p})$ for $a_n(H) = a_n(H, p)$.

Hence, in view of Theorem 1, we have to check, whether for all $H \in \mathcal{H}_0$, we have that $n\widehat{p}^{r(H)} \xrightarrow[n\to\infty]{} \infty$ in probability, the coefficients

$$\widehat{b}(H) = \sup_n \frac{n^{n(H)/2} \widehat{p}^{m(H)/2} a_n^*(H)}{\sqrt{\mathbb{V}\mathrm{ar}^*(X_n^*)}} \tag{4.22}$$

are bounded in probability and satisfy

$$\sum_{H \in \mathcal{H}_0} \widehat{b}^2(H) |\mathrm{aut}(H)| = O_P(1), \tag{4.23}$$

respectively. For this purpose, let us consider $\widehat{p} = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} A_{ij}$ in more detail. To be more precise, as p_{ij}, and consequently also the distribution of A_{ij}, are allowed to depend on n, we will write $A_{n,ij}$ and $p_{n,ij}$ in the following. For the expectation, we get

$$\mathbb{E}(\widehat{p}) = \mathbb{E}\left(\binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} A_{n,ij}\right) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} \mathbb{E}(A_{n,ij}) = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} p_{n,ij} = p_{\mathrm{mean}}.$$

For the variance, as the edges are formed independently according to Bernoulli distributions with connection probabilities $p_{n,ij}$, we have

$$\mathbb{V}\mathrm{ar}(\widehat{p}) = \binom{n}{2}^{-2} \sum_{1\leq i<j\leq n} \mathbb{V}\mathrm{ar}(A_{n,ij}) = \binom{n}{2}^{-2} \sum_{1\leq i<j\leq n} p_{n,ij}(1-p_{n,ij}).$$

Hence, this leads to

$$\widehat{p} - p_{\mathrm{mean}} = O_P\left(\sqrt{\binom{n}{2}^{-2} \sum_{1\leq i<j\leq n} p_{n,ij}(1-p_{n,ij})}\right) \leq O_P\left(\frac{\min\{p_{\mathrm{mean}}, 1-p_{\mathrm{mean}}\}}{n}\right) \leq O_P\left(\frac{1}{n}\right) \quad (4.24)$$

due to $\binom{n}{2} = n^2$ and $1 - p_{n,ij} \leq 1$ and $\sup_{i,j} p_{ij} = o(1)$ if $p_0 = 0$ and $\sup_{n;i,j} p_{ij} < 1$ and $\inf_{n;i,j} p_{ij} > 0$ if $p_0 \in (0,1)$. Then, a Taylor series argument leads to

$$\begin{aligned} np^{r(H)} &= n\left(p^{r(H)} + r(H)\widetilde{p}^{r(H)-1}(\widehat{p}-p)\right) = np^{r(H)} + O_P\left(nr(h)\widetilde{p}^{r(H)-1}\frac{\sqrt{p}}{n}\right) \\ &= np^{r(H)} + O_P\left(n\left(\frac{\sqrt{p}}{n}\right)^{r(H)-1}\frac{\sqrt{p}}{n}\right) = np^{r(H)} + O_P\left(n\left(\frac{\sqrt{p}}{n}\right)^{r(H)}\right) \\ &\to \infty \end{aligned}$$

in probability, because $r(h) \geq 1$. To show that (4.22) is bounded in probability, note that $X_n^* = \sum_{H\in\mathcal{H}_0} a_n^*(H)S_n^*(H)$ with $a_n^*(H) = a_n(H,\widehat{p})$ and

$$\mathbb{V}\mathrm{ar}^*(X_n^*) = \sum_{H\in\mathcal{H}_0} a_n^{*2}(H)\mathbb{V}\mathrm{ar}(S_n^*(H)),$$

where, according to Proposition 1,

$$\mathbb{V}\mathrm{ar}^*(S_n^*(H)) = |\mathrm{aut}(H)|(n)_{n(H)}(\widehat{p}(1-\widehat{p}))^{m(H)}.$$

By plugging-in, we get

$$\widehat{b}(H) = \sup_n \frac{n^{n(H)/2}\widehat{p}^{m(H)/2}a_n(H)}{\sqrt{\mathbb{V}\mathrm{ar}^*(X_n^*)}} = \sup_n \frac{n^{n(H)/2}\widehat{p}^{m(H)/2}a_n(H)}{\sqrt{|\mathrm{aut}(H)|(n)_{n(H)}(\widehat{p}(1-\widehat{p}))^{m(H)}}}$$

which remains bounded in probability by similar arguments as above due to (4.16) and making use of (4.24). Finally, using the same arguments, (4.23) follows from (4.17). □

Appendix B: Additional tables and figures

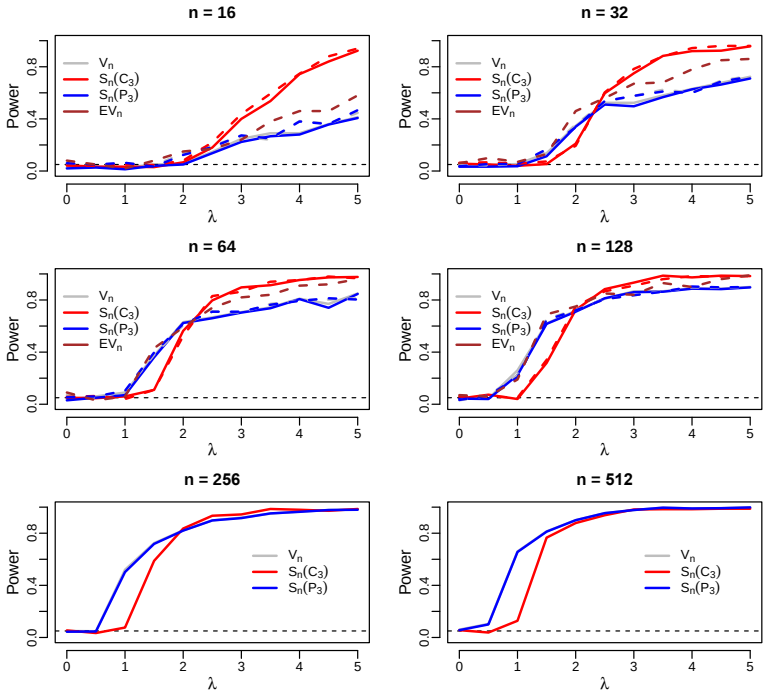


Figure 20: Power of the tests V_n, E_n, $S_n(C_3)$ and $S_n(P_3)$ for the alternative of a SBM setup with three blocks with $p_{\text{mean}} = \frac{\log n}{\sqrt{n}}$. The asymptotic versions of the tests are represented by solid lines and the bootstrap version by dashed lines.

5

Dynamic change detection in topics based on rolling LDAs

Based on: Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf & Carsten Jentsch (2022): "Dynamic change detection in topics based on rolling LDAs", In: *Proceedings in the Text2Story'22 Workshop*. Vol. 3117.CEUR-WS, pp. 5–13.

Abstract

Topic modeling methods such as e.g. Latent Dirichlet Allocation (LDA) are popular techniques to analyze large text corpora. With huge amounts of textual data that are collected over time in various fields of applied research, it becomes also relevant to be able to automatically monitor the evolution of topics identified from some sort of dynamic topic modeling approach. For this purpose, we propose a dynamic change detection method that relies on a rolling version of the classical LDA that allows for coherently modeled topics over time that are able to adapt to changing vocabulary. The changes are detected by assessing the intensity of word change in the LDA's topics over time in comparison to the expected intensity of word change under stable conditions using resampling techniques. We apply our method to topics obtained by applying the RollingLDA to Covid-19 related news data from CNN and illustrate that the detected changes in these topics are well interpretable.

5.1 Introduction

While change detection is an active field in modern research, the application for text data poses even further obstacles due to its unstructured nature. And yet, an effective method for change detection would have many use cases. Particularly, when dealing with large text

corpora collected over time, an online detection approach will be useful to analyze the evolution of narratives or to spot a shift in a discourse about certain topics. For this purpose, we propose an online change detection method for text data by analyzing the change of word distributions within topics of Latent Dirichlet Allocation (LDA) models. As we are dealing with time series of textual data, we make use of a rolling version of the classical LDA, called RollingLDA [Rieger et al., 2021]. The method is designed to construct coherently interpretable topics modeled over time that are allowed to adapt to a changing vocabulary. The changes are detected by dynamically assessing the change intensity in word usage in the LDA's topics over time in comparison to the change intensity expected in stable periods using resampling techniques.

The main goal of change detection is to identify possible anomalies in a process. Typically, there are the two perspectives towards this issue: offline and online applications. Our approach is applicable for both tasks, but, for each time point, it relies exclusively on the text data that has already been observed. Hence, we focus on the usually more relevant task of online monitoring. In traditional schemes for change detection [Montgomery, 2007; Oakland and Oakland, 2018], control charts are applied to visualize the monitoring procedure using a control statistic which is successively calculated for each time point. An alarm is triggered whenever the statistic lies outside of some control limits. In practice, there are a variety of different control charts including memory-free setups (e.g. Shewhart charts) and memory-based charts (e.g. EWMA, CUSUM), However, these traditional procedures can not be applied to textual data off the shelf because of the high dimensionality of large text corpora. In addition, an in-control state to reliably calculate the control limits is frequently not available due to the strong dynamics in text data, e.g. newspaper articles. To overcome these issues, we propose to use a control statistic based on a similarity metric that represents the resemblance of topic's word distributions over consecutive time points. Control limits are derived by a resampling procedure using word count vectors based on time-variant topics modeled by RollingLDA.

In a similar context, the usage of LDA was proposed for change point detection for topic distributions in texts [Bose and Mukherjee, 2021], which is based on a modified version of the wild binary segmentation algorithm [Fryzlewicz, 2014] designed for offline detection setups. There is also work considering Bayesian online monitoring [Adams and MacKay, 2007] for textual data using a document-based model [Kim and Choi, 2015] and an approach based on similarity metrics, which aims to detect global events in topics in offline settings [Keane et al., 2015]. There is also work which analyzes the transitions of narratives between topics [Mei and Zhai, 2005]. In contrast, the rolling window approach of RollingLDA constructs coherently interpretable topics modeled over time and allows the resulting dynamic change detection method to become applicable in online settings. Compared to the mentioned related methods, our method is designed to detect changes in word distributions of topics over time rather than global changes in topic distributions (of sets) of documents e.g. Bose and Mukherjee, 2021; Keane et al., 2015; Kim and Choi, 2015 or sentiments in topics e.g. Liang and Wang, 2019 or

(in contrast) changes in topic distributions of words e.g. Frermann and Lapata, 2016. This results in a more refined monitoring procedure that allows for the detection of narrative shifts that are changing the word usage within a certain topic instead of measuring the frequency of a topic over time within the whole corpus. Building on this, we aim that our proposed method can provide groundwork for the extraction and temporal localization of narratives in texts.

5.2 Methodological framework

For the proposed change detection algorithm, we make use of the existing method of a rolling version of the classical LDA (RollingLDA) to construct coherent topics over time and measure similarities of topics for consecutive time points using the well-established cosine similarity.

5.2.1 Latent Dirichlet Allocation

The classical LDA [Blei et al., 2003b] models distributions of K latent topics for each text. Let $W_n^{(m)}$ be a single word token at position $n = 1, \dots N^{(m)}$ in text $m = 1, \dots, M$ of a corpus of M texts. Then, a single text is given by

$$\boldsymbol{D}^{(m)} = \left(W_1^{(m)}, \dots, W_{N^{(m)}}^{(m)}\right), \quad W_n^{(m)} \in \boldsymbol{W} = \{W_1, \dots, W_V\}, V = |\boldsymbol{W}|$$

and the corresponding topic assignments for each text are given by

$$\boldsymbol{T}^{(m)} = \left(T_1^{(m)}, \dots, T_{N^{(m)}}^{(m)}\right), \quad T_n^{(m)} \in \boldsymbol{T} = \{T_1, \dots, T_K\}.$$

From this, let $n_k^{(mv)}, k = 1, \dots, K, v = 1, \dots, V$ denote the number of assignments of word v in text m to topic k. Then, we define the cumulative count of word v in topic k over all texts by $n_k^{(\bullet v)}$ and denote the total count of assignments to topic k by $n_k^{(\bullet\bullet)}$. Using these definitions, the underlying probability model [Griffiths and Steyvers, 2004] can be written as

$$W_n^{(m)} \mid T_n^{(m)}, \phi_k \sim \text{Discr}(\phi_k), \quad \phi_k \sim \text{Dir}(\eta), \quad T_n^{(m)} \mid \theta_m \sim \text{Discr}(\theta_m), \quad \theta_m \sim \text{Dir}(\alpha).$$

For a given parameter set $\{K, \alpha, \eta\}$, with the Dirichlet priors α and η defining the type of mixture of topics in every text and the type of mixture of words in every topic, LDA assigns one of the K topics to each token. A word distribution estimator per topic for $\phi_k = (\phi_{k,1}, \dots, \phi_{k,V})^T \in (0,1)^V$ can be derived through the collapsed Gibbs sampler procedure [Griffiths and Steyvers, 2004] by

$$\hat{\phi}_{k,v} = \frac{n_k^{(\bullet v)} + \eta}{n_k^{(\bullet\bullet)} + V\eta}. \tag{5.25}$$

5.2.2 RollingLDA

RollingLDA [Rieger et al., 2021] is a rolling version of classical LDA. New texts are modeled based on existing topics of the previous model. Thereby, not the whole knowledge of the entire past of the model is used, but only the information of the topics from more recent texts based on a user-chosen memory parameter. For each time point, based on the topic assignments within this memory period, the topics are initialized and modeled forward. This form of modeling preserves the topic structure of the model so that topics remain coherently interpretable over time. At the same time, constraining the knowledge of the model to the user-chosen memory period allows for changes in topics based on new vocabulary or word choices. There are other dynamic variants of the LDA approach [Blei et al., 2003a; Blei and Lafferty, 2006; Song et al., 2005; Wang et al., 2008; Wang and McCallum, 2006] deliberately designed to model gradual changes, and therefore not as well suited to detect abrupt changes. We use the update algorithm RollingLDA to make our proposed change detection method applicable in an online manner. Thereby, a text is assigned to a time point on the basis of its publication date. The step size of the and is chosen on a weekly basis in the present case as this seems natural for journalistic texts.

5.2.3 Similarity

Our change detection algorithm builds on a similarity measure for word count vectors. Following up on the notation from Section 5.2.1 the word count vector for topic $k \in \{1, \ldots, K\}$ at one time point $t \in \{0, \ldots, T\}$ is given by

$$\boldsymbol{n}_{k|t} = \left(n_{k|t}^{(\bullet 1)}, \ldots, n_{k|t}^{(\bullet V)}\right)^T \in \mathbb{N}_0^V = \{0, 1, 2, \ldots\}^V.$$

Then, monitoring the similarity of topics over time for (consecutive) time points t_1 and t_2 is done using the cosine similarity

$$\cos\left(\boldsymbol{n}_{k|t_1}, \boldsymbol{n}_{k|t_2}\right) = \frac{\sum_v n_{k|t_1}^{(\bullet v)} n_{k|t_2}^{(\bullet v)}}{\sqrt{\sum_v \left(n_{k|t_1}^{(\bullet v)}\right)^2} \sqrt{\sum_v \left(n_{k|t_2}^{(\bullet v)}\right)^2}}.$$

The choice of cosine similarity is common in the context of change point detection for text data e.g. Keane et al., 2015; Wang and Goutte, 2018. Compared to other similarity measures such as the Jaccard coefficient, Jensen-Shannon Divergence, χ^2-, Hellinger and Manhattan Distance, the cosine similarity fulfills some typical conditions required for monitoring a similarity measure [Rieger et al., 2021].

5.3 Change detection

In combination with the existing method RollingLDA and cosine similarity, our contributed method for change detection relies on classical resampling approaches to identify changes within topics. We estimate the realized change in a topic based on the similarity between the current and previous count vectors of word assignments and compare the resulting similarity score to resampling-based similarity scores which are generated under stable conditions, such that no extraordinary changes occurred in the topic.

5.3.1 Set of changes

Suppose we consider K topics over T time points to be monitored. If the actual observed similarity of the word vector of some topic $k \in \{1, \dots, K\}$ at some time $t \in \{0, 1, \dots, T\}$ given by $\boldsymbol{n}_{k|t}$, compared to the frequency vector of the topic over a predefined reference time period $t - z_k^t, \dots, t-1$, given by

$$\boldsymbol{n}_{k|(t-z_k^t):(t-1)} = \sum_{z=1}^{z_k^t} \boldsymbol{n}_{k|t-z},$$

is smaller than a threshold q_k^t which is calibrated based on similarities under stable conditions (see Section 5.3.2), then we identify a change within topic k at time t. The set of identified changes in topic k up to time point t can then be defined as

$$C_k^t = \left\{ u \mid 0 < u \leq t \leq T : \cos\left(\boldsymbol{n}_{k|u}, \boldsymbol{n}_{k|(u-z_k^u):(u-1)}\right) < q_k^t \right\} \cup 0, \tag{5.26}$$

where the time point $t = 0$ is always included for technical reasons, to compute the current run length without a change $z_k^t = \min\left\{z_{\max}, t - \max C_k^{t-1}\right\}$. Thus, the reference period spans the last $z_{\max}$ time points if no change was detected during that time, and spans the time that has passed since the last change, otherwise. The parameter $z_{\max}$ is to be chosen by the user and is intended to smooth the similarities to prevent from detecting false positives.

5.3.2 Dynamic thresholds

For the calculation of the threshold q_k^t, the estimated word distribution of a topic k at some time point t, as well as over the corresponding reference period $t - z_k^t, \dots, t-1$ are needed. For this, let $\hat{\phi}_k^t$ and $\hat{\phi}_k^{(t-z_k^t):(t-1)}$ be defined by

$$\hat{\phi}_{k,v}^t = \frac{n_{k|t}^{(\bullet v)} + \eta}{n_{k|t}^{(\bullet\bullet)} + V\eta} \quad \text{and} \quad \hat{\phi}_{k,v}^{(t-z_k^t):(t-1)} = \frac{n_{k|(t-z_k^t):(t-1)}^{(\bullet v)} + \eta}{n_{k|(t-z_k^t):(t-1)}^{(\bullet\bullet)} + V\eta}$$

analogously to Equation (5.25).

The application of the change point detection algorithm is designed for text data, more precisely for empirical word distributions of K topics modeled by LDA in a given text corpus. Since word choice - especially in journalistic texts - varies considerably over time, a situation in which there is no change in the word distribution within topics across consecutive time points does not reflect the expected situation. Rather, it is to be expected that topics change gradually on an ongoing basis. Accordingly, our method aims to identify not the numerous customary changes in the topics, but the unexpectedly large ones. To do so, we define an expected word distribution $\tilde{\phi}_k^{(t)}$ for time point t under stable conditions that include the customary changes as a convex combination of the two estimators of the word distribution of topic k, one for the reference time period $t - z_k^t, \ldots, t-1$ and one for the current time point t. Using the mixture parameter $p \in [0, 1]$, which can be tuned based on how substantial the detected changes should be, the intensity of the expected change is considered in the determination of this estimator by

$$\tilde{\phi}_k^{(t)} = (1-p)\, \hat{\phi}_{k,v}^{(t-z_k^t):(t-1)} + p\, \hat{\phi}_{k,v}^{(t)}.$$

Our method uses the estimator $\tilde{\phi}_k^{(t)}$ to simulate R expected word count vectors $\tilde{\boldsymbol{n}}_{k|t}^r, r = 1, \ldots, R$ based on a parametric bootstrap approach. In this process, each word is drawn according to its estimated probability of occurrence regarding $\tilde{\phi}_k^{(t)}$ and each sample r consists of $n_{k|t}^{(\bullet\bullet)}$ draws, the number of words assigned to topic k at time point t. Then, we calculate the cosine similarity

$$\cos\left(\tilde{\boldsymbol{n}}_{k|t}^r, \boldsymbol{n}_{k|(t-z_k^t):(t-1)}\right)$$

for each of the $r = 1, \ldots, R$ bootstrap samples and set the threshold q_k^t equal to the 0.01 quantile of these simulated similarity values generated under stable conditions. Combinations of topics and time points for which the observed similarity is smaller than the corresponding quantile are classified as change points according to Equation (5.26).

5.4 Analysis

For conducting the real data analysis, the data set under study was created with Python, whereas the preprocessing, the modeling, all postprocessing steps and analyses are performed using R. The scripts for all analysis steps can be found in the associated GitHub repository github.com/JonasRieger/topicalchanges.

5.4.1 Data and study design

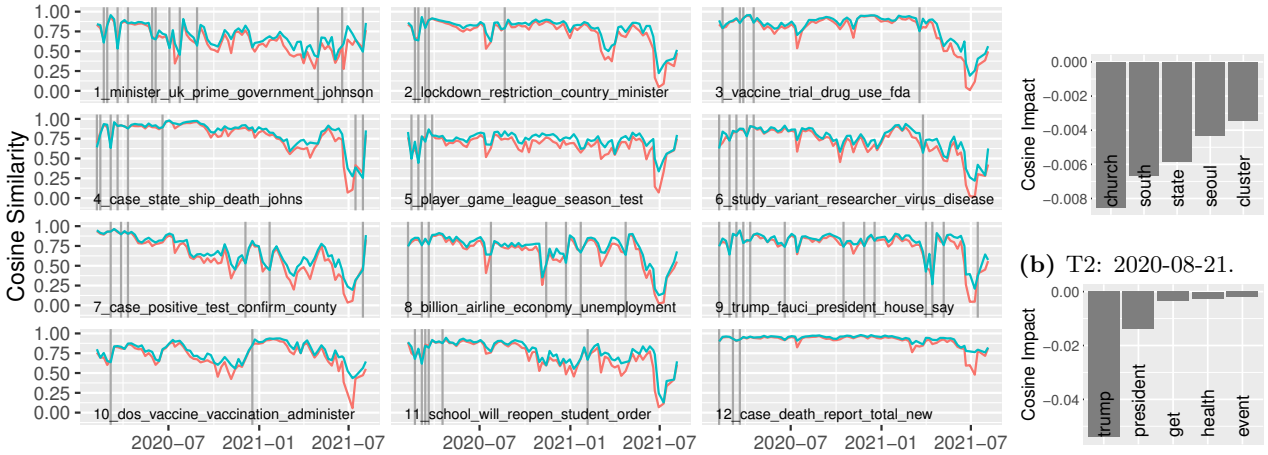
To assess the quality of our change point algorithm, we use the TLS-Covid19 data set [Pasquali et al., 2021]. It is generated using Covid-19 related liveblog articles of CNN, collected from January 22nd 2020 up until December 12th 2021. Each liveblog is interpreted to belong to a topic and comprises texts and key moments. The texts form a time line containing events, which are summarized by its key moments. The resulting corpus consists of 27,432 texts and 1,462 key moments. Although the data set contains multiple key moments per day on average, we do not consider all them a change point as our aim is to detect larger changes based on aggregated weekly texts. However, these key moments serve well as indicators, which enable us to check whether the detected changes are actually related to real events or if they are false positives.

We use common NLP preprocessing steps for the texts, i.e. characters are formatted to lowercase, numbers and punctuation are removed. Moreover, a trusted stopword list is applied to remove words that do not help in classifying texts in topics, we use a lemmatization dictionary (github.com/michmech/lemmatization-lists) and neglect words with less than two characters.

We model the CNN data set using RollingLDA on a weekly basis, starting on Saturday of each week, and we consider the previous week as initialization for the model's topics. The first 10 days of modeling, Wednesday, January 22nd 2020 until Friday, January 31st 2020, serve as the initial chunk corresponding to $t = 0$. During this period, 605 texts were published. In the data set, there are weeks that do not contain any texts. In this case, the corresponding time point is omitted. Then, to model the texts of the following chunk, at least the last 10 texts are used, as well as all other texts published on the same date as the oldest of these 10 texts. As parameters, we assume $K = 12$ topics, define the reference period of the topics to the last $z_{\max} = 4$ weeks, and choose $p = 0.85$, since these values are accountable by plausibility and seem to yield reasonable results. For other parameter choices, i.e. $K = 8, \ldots, 20$, $z_{\max} = 1, \ldots, 20$, $p = 0.5, \ldots 0.8, 0.81, \ldots, 0.90$, results can be found in our associated repository.

5.4.2 Findings

The results of our chosen model are displayed in Figure 21. Figure 21a shows the detected changes by vertical gray lines, which are the weeks in which the observed similarity (blue curve) is lower than the expected one (red curve). Furthermore, for two changes we show which words are mainly causing the detection of the change. The score of a word in a topic at a given time point is calculated by the topic's similarity without considering this word and subtracting it from the actual realized similarity. These leave-one-out cosine impact scores for the words with the five most negative scores are shown in Figure 21b and 21c. In general, most of the



(a) Similarities (blue), thresholds (red) and detected changes (vertical lines, gray).

(b) T2: 2020-08-21.

(c) T9: 2020-10-16.

Figure 21: Similarity values, thresholds, and detected changes over the observation period for all $K = 12$ topics, as well as the five most influential words for two selected change points in topics 2 and 9.

changes we detect occur within the first four months of 2020. This is because the wording was constantly changing, as the Covid-19 epidemic turned into a pandemic over the course of these months. New people and organizations were associated with Covid-19, which is why we detect a bunch of consecutive changes in every topic. As the pandemic reached out into further countries, the detected changes became less frequent for most topics. In the following we share our interpretation of some exemplary detected changes.

The third topic, containing information about vaccination and testing procedures, shows a change in the week starting on the 13th of March 2021. In this week, the AstraZeneca vaccination process in several EU-states was stopped due the risk of causing blood clots.[1] The sixth topic, a topic about medical studies and research, shows a change in the following week, in which AstraZeneca presented a study about the effectiveness of its vaccine.[2] Another interesting detection is the change in the vaccination-related topic 10 in December 2020, just as the vaccination process started in the US.[3]

Political changes are also detected in several topics, such as the start of Joe Biden's presidential era in late January 2021 in topic 11, the return of Donald Trump to office after his Covid-19

[1] CNN online, 2021-03-15 3:03 p.m. ET, "Spain joins Germany, France and Italy in halting AstraZeneca Covid-19 vaccinations", `https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-03-15-21/h_d938057f2ef588f74565bdbb01f12387`, visited on 2022-01-20.

[2] CNN online, 2021-03-25 2:48 a.m. ET, "New AstraZeneca report says vaccine was 76% effective in preventing Covid-19 symptoms", `https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-03-25-21/h_9f01e2e53b62873f1c742254d27fbf5f`, visited on 2022-01-20.

[3] CNN online, 2020-12-14 10:08 p.m. ET, "The first doses of FDA-authorized Covid-19 vaccine were administered in the US. Here's what we know", `https://edition.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-12-15-20/h_32be1a72dc05f874eda167c95c8f1bba`, visited on 2022-01-20.

infection in October 2020[4] in topic 9 (cf. Figure 21c) or the discussion about the origin of the virus after a WHO report in late March 2021 in topic 9.[5] A Covid-19 outbreak in the South Korean Sarang-jeil church in August 2020[6] is detected in topic 2 (cf. Figure 21b).

While these topics detect changes across the entire time span, the twelfth topic, representing the report of the current number of Covid cases, does not detect a single change after March 2020. This is most likely because, after the pandemic had reached the US and Europe in early 2020, the number of cases was consistently reported and the interpretations and implication of those case numbers are detected as changes in other topics. Even in the last months of the data set, in which the number of texts decreased and the results thus show a lower similarity, the twelfth topic retained a rather high similarity of above 0.75.

5.5 Discussion

In this paper, we presented a novel change detection method for text data. To construct coherently interpretably topics, we used RollingLDA to model a time series on textual data and compared the model's word distribution vectors with those of texts resampled under stable conditions. We applied our model on the TLS-Covid19 data set consisting of Covid-19 related news articles from CNN between January 2020 and December 2021.

Our method detects several meaningful changes in the evolving news coverage during the pandemic, including e.g. the start of vaccinations and several controversies over the course of the vaccination campaign as well as political changes such as the start of Joe Biden's presidential era. Out of 78 detected changes, we were instantly able to judge 55 (71%) as plausible ones based on manual labeling using the leave-one-out cosine impacts (cf. Figure 21b, 21c and repository). The share increases to 78% if we exclude the turbulent initial phase of the Covid-19 pandemic and only consider changes since April 2020. While we cannot tell how many changes were missed out that could be considered as important as the ones mentioned above, our model contains a mixture parameter to calibrate the detection for general change of topics within a usual news week. If more, but less substantial or less, but more substantial changes are to be detected, this parameter p can be tuned accordingly. In combination with the maximum length of the reference period $z_{\max}$, the set $\{p, z_{\max}\}$ forms the model's hyperparameters to be optimized.

[4] CNN online, 2020-10-12 12:01 a.m. ET, "Trump says he tested 'totally negative' for Covid-19", `https://edition.cnn.com/world/live-news/coronavirus-pandemic-10-12-20-intl/h_7570d53b184a5b1d6ec97ce67330e4c9`, visited on 2022-01-20.

[5] CNN online, 2021-03-29 11:22 a.m. ET, "Upcoming WHO report will deem Covid-19 lab leak extremely unlikely, source says", `https://www.cnn.com/world/live-news/coronavirus-pandemic-vaccine-updates-03-29-21/h_1f239fee1b0584ca9a5b6085357ac907`, visited on 2022-01-20.

[6] CNN online, 2020-08-20 12:55 a.m. ET, "South Korea's latest church-linked coronavirus outbreak is turning into a battle over religious freedom", `https://edition.cnn.com/world/live-news/coronavirus-pandemic-08-20-20-intl/h_288a15acd1b29e732c4e10693641088a`, visited on 2022-01-20.

6

Visually analyzing topic change points in temporal text collections

Based on: Krause, Cedric, Jonas Rieger, Jonathan Flossdorf, Carsten Jentsch, and Fabian Beck (2023). "Visually Analyzing Topic Change Points in Temporal Text Collections". In: *Vision, Modeling, and Visualization*. Ed. by Michael Guthe and Thorsten Grosch. The Eurographics Association. ISBN: 978-3-03868-232-5. DOI: 10.2312/vmv.20231231.

Abstract

Texts are collected over time to document important news and facts on different topics, and they reflect temporal changes in the concepts that they cover. While some changes might slowly evolve, other changes abruptly surface as explicit change points. In an application study for a specific change point extraction method based on a rolling Latent Dirichlet Allocation (LDA), we have developed a visualization approach that allows exploring such change points and related change patterns. In contrast to most previous approaches, we investigate the visual characterization of change points within such topics in greater depth and target the approach at statistical experts. Our visualizations not only provides an overview of topics, but supports the detailed exploration of temporal developments. The interplay of general topic contents, development, and similarities with detected change points reveals rich insights into different kinds of change patterns. The approach comprises a combination of views including topic timeline representations with detected change points, comparative word clouds, and temporal similarity matrices. In an interactive exploration, these views adapt to selected topics, words, or points in time. We demonstrate in an in-depth application example involving statistical experts that the use cases of our approach are twofold: first, supporting the experts in a qualitative evaluation and fine-tuning of their model and, second, providing them insights that are relevant to end users of their model.

6.1 Introduction

In large text corpora, like news collections, diverse stories are covered. Topic extraction methods can automatically identify topics by grouping related stories. This is helpful in various scenarios, from the identification of articles of interest for a person, to the support for a general overview of the text collection. However, whereas the topics are often treated as static concepts, in practice, they are dynamic. With new texts being added over time, the coverage of the derived topics changes. Analyzing current media coverage incrementally, algorithmic methods employ statistical indicators to detect change points at which a topic is sufficiently different from its previous state. However, these methods often appear as black-boxes as they do not explain these changes. In order to understand the result, it is necessary to analyze all output artifacts of the models. Inspecting the contents of the derived topics, their similarities with each other, and change points within the topics that occurred over time can reveal rich insights about the underlying data. And even more, it may help developers of such models to inform potential adjustments to the analysis method or its parameter choices.

As a team of statistics and visualization researchers, we co-developed an approach to enhance the visual analysis of such models' results. To this end, we have identified relevant analysis questions, that are necessary to reason about the quality of the topic extraction and change detection. Based on the analysis questions, we designed a conceptual approach to address these questions, implemented a corresponding visualization system, and classified common change patterns that we identified while using our system.

Our visualization system in Figure 22 shows a word cloud that adapts word sizes and other encodings to the selection topics and time steps. Each topic is individually represented in a timeline visualization, including detected change points that mark abrupt changes within the vocabulary of the topic. A butterfly bar chart in the change detail view displays which words in the dynamic vocabulary changed most noticeably and caused the detection of a change point. Moreover, we use similarity matrices to visualize intra-topic similarity, as well as pairwise similarity between different topics over time.

In this work, we use a rolling version of Latent Dirichlet Allocation (LDA) [Rieger et al., 2021] as a dynamic topic extraction method on a CNN news collection about COVID-19 from 2020 and 2021. We apply change point detection [Rieger et al., 2022] to the derived topics to statistically identify abrupt changes in the topics' vocabularies. Originally, the method is designed as an online analysis method to detect changes right when they happen to gather insights about the data. However, since our main target users are the model developers, we look at the results retrospectively over a longer time period to enable qualitative evaluation of whether the detection works as intended. We show that our *vis-for-stats* approach helps to

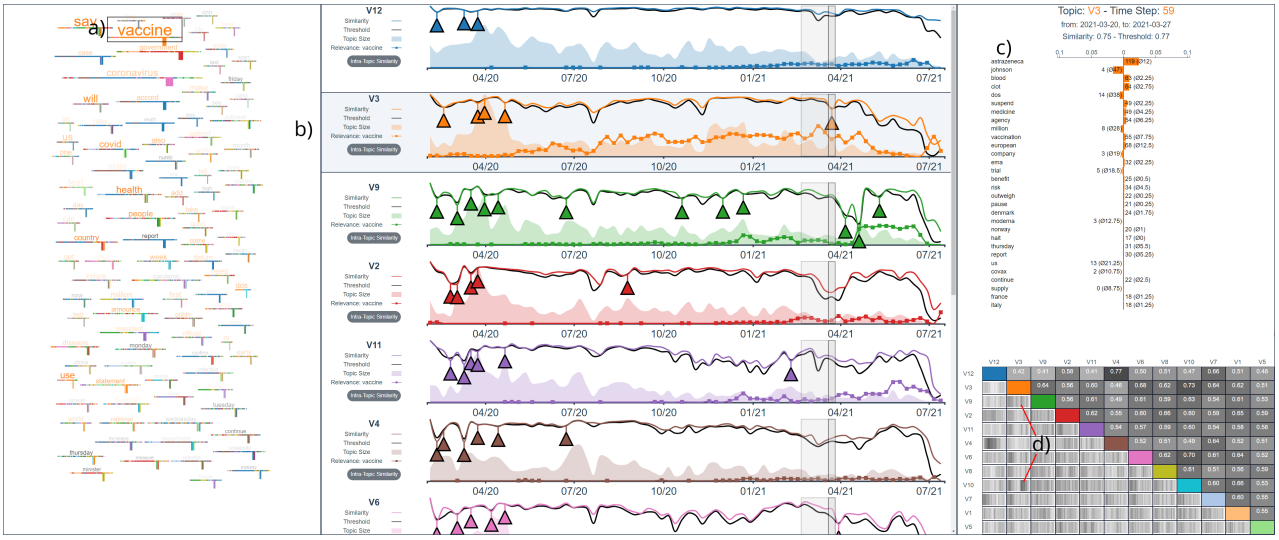


Figure 22: Our visualization system displays a dynamic topic extraction from a COVID-19-related news dataset. The selected word *'vaccine'* (a) is especially important in the selected topic V3 (b). A triangle marks a detected change point in the topic around March of 2021, when the word *'astrazeneca'* (c) gained substantial importance, compared to the reference period. This change point is detected at a time when the topic possesses a comparatively high similarity with topics V9 and V10 (d).

answer the identified analysis questions and, thus, supports experts in model calibration and helps in identifying insights relevant to end users of the model.

Our application study, hence, suggests a tailored visualization solution for a specific statistical approach, targeting experts as users. This limits the immediate applicability of our approach. However, our contributions span broader, as they also include results transferable to related problems and interfaces:

(i) analysis questions to analyze temporal text collections along topics and detected change points,

(ii) reusable concepts for the visual analysis of dynamic text collections from overview, over topic evolution, to detailed inspection of change points,

(iii) a classification of observed change patterns based on the characteristics of change points, and

(iv) exemplary insights from an extensive application example with real world data.

6.2 Related Work

We are not the first to visualize changing topics in document collections; approaches are numerous and constitute separate sections in literature surveys on visual text analysis [Dou

and Liu, 2016] and visualization of scientific literature topics [Zhang et al., 2017]. For instance, the well-known *ThemeRiver* [Havre et al., 2002] approach visualizes changing topic sizes as a variant of stacked area charts on a timeline. Other approaches adapt this metaphor and, for instance, extend it to branching and merging streams [Cui et al., 2011, 2014; Liu et al., 2016; Xu et al., 2013]. But also small multiple representations are common, where different topics are shown as different rows on the same timeline [Dou et al., 2012; Krstajic et al., 2011; Luo et al., 2012]. We decided to use such small multiples, and not stacked streams, for the topics because our scenario did not allow computing quantifiable exchange of content between topics and stacking streams would have generally limited the options to encode additional information within the topic representations.

Also, many approaches have considered specific events that mark changes in the dynamic topics. Within branching and merging streams, the branch and merge points constitute a form of discrete event. These are detected, for instance, by incrementally updating topic models and applying defined branch and merge criteria [Cui et al., 2011; Gad et al., 2015] or evolutionary hierarchical topic trees combined with a dynamic tree cut method [Cui et al., 2014; Liu et al., 2016]. Some change events are specialized to certain applications, for instance, whether a topic changes from cooperation to competition within a debate [Sun et al., 2014]. Other techniques consider events within consistent topics. These can be the real world events derived from keywords that were used within a topic's documents. The emergence of coverage of that event can also be considered a change point in the topic. *EventRiver* [Luo et al., 2012] visualizes these events as drops that encode the number of articles that are closely related content-wise and time-wise. Lu et al., 2018 follow a similar procedure, but allow the user to handpick events to annotate the topic for a more user-guided analysis. *LeadLine* [Dou et al., 2012], *TopTom* [Gobbo et al., 2019], and *TwitInfo* [Marcus et al., 2011] focus on change points within consistent topics. Using statistical methods, they identify peaks and valleys in the number of articles covering the topic. In contrast, we focus on structural breaks in the statistical analysis of the vocabulary used in the topic. We are not aware of any approach with similar focus. Moreover, we observe that the discussed approaches seem to rather target end users from an application domain (e.g., journalists or historians) and, hence, try to abstract from technical details. However, our approach is made for statistical experts, who would also be interested in such details and use them for evaluating and adapting the detection method.

While we employ and adapt various standard visualization techniques in our approach, our use of word clouds might be most specialized. For our interactive, topic- and time-aware word cloud, we drew inspiration from several previous works. Time-aware word clouds, like *PyramidTags* [Knittel et al., 2021], have been proposed help analysts explore dynamic document collections. Together with topic extraction, they have been utilized to either display relevant words on the topic streams themselves [Liu et al., 2012] or can be accessed through an on-demand lens [Xu et al., 2013]. The word cloud visualization we use is also adaptive to the

topics and time steps, but keeps a global layout, to enable comparison of the words across time and topic boundaries. *RadCloud* [Burch et al., 2014] uses stacked bar charts below the words to encode their relevance in multiple categories. We adopt this idea and use scarfplots in our word cloud to attribute words to topics in a glance.

6.3 Analysis Questions

Agnostic of the concrete methods used for topic extraction and change point detection, we formulate analysis questions to reason about the quality of such models. Based on our collaboration between visualization and statistics researchers, we have tried to capture those questions statistical experts would explicitly or implicitly want to answer when working with a visualization approach to analyze their methods. While initial questions were formulated early and guided the design of the system, we have refined and consolidated the questions iteratively throughout the process. In the paper, we use them as a reference to connect the concepts and results discussed in different sections.

We start from a collection of documents and extract a set of *topics* at different points in time. Each derived topic may reflect multiple real-world *concerns* (e.g., a person, an entity, a concept), and it is crucial to understand what the topics capture exactly. The general *importance* of a topic may vary over time, as well as the concerns that are connected. *Similarity* between topics is expressed by the partial overlap or inherent semantic connections between their concerns. Understanding this similarity and also identifying if concerns shift between the topics provides important context to make sense of the temporal evolution of the topics.

> ***AQ 1*** *– Topic Evolution*
>
> **AQ 1.1** What are the importance and concerns of a topic (over time)?
>
> **AQ 1.2** How similar are topics to each other (over time)?
>
> **AQ 1.3** Do concerns move from one topic to another?

The temporal development of the extracted topics can be accompanied by a number of *change points* at which its deviation from previous states is declared as substantial. Causes for such deviations are typically new developments and events that lead to adaptions in the concerns covered by the topic, and they can potentially impact multiple topics. Each change point is further characterized by its context within and beyond the topic (e.g., the topic was not important at the time of the change).

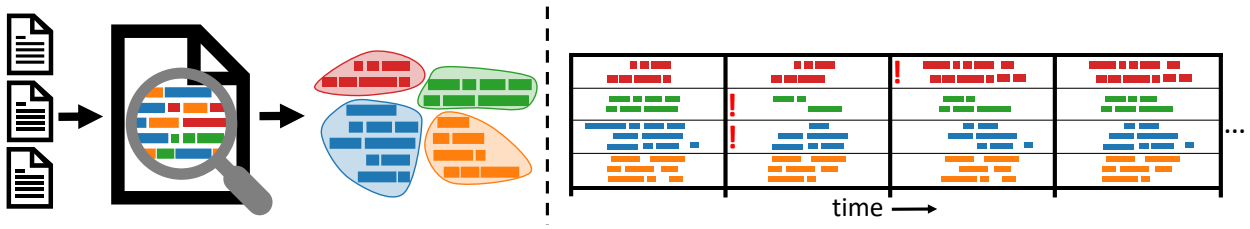


Figure 23: LDA takes a text collection and allocates the words in each document to a predefined number of topics. Each word can only belong to one topic, but several instances of the same word (e.g., in different documents) can be assigned to different topics. Topics, therefore, are represented by sets of words (left). Repeating the allocation, on rolling time windows, results in dynamic vocabularies. If the vocabulary of a topic at a given time step is sufficiently different from previous states, the change point detection method leaves a mark (right).

***AQ 2** – Change Points*

AQ 2.1 When do change points occur and what are their characteristics?

AQ 2.2 What shifts in the concerns caused a change point?

AQ 2.3 Do multiple change points share characteristics and causes?

The grouping into *Topic Evolution* and *Change Points* structures the questions according to the two abstraction steps made (*documents* → *topics* and *topics* → *change points*), but does not divide the analysis. The questions may contain connections across the boundaries of the grouping. Observing a peculiar instance in either, might initiate a new investigation in the other block. For instance, an analysis of specific change points can lead to questions of the general importance of a topic and finding related topics.

6.4 Application Scenario

Whereas the analysis questions are formulated independent of a concrete statistical method, we focus our application study on a specific scenario. On the one hand, this is necessary to operationalize, for instance, how *concerns* are described that reflect in a topic or how *similarities* of topics can be computed. On the other hand, we want to provide deep insights for statistical experts that take into account the specifics of a method. Hence, our approach can be considered a *white-box analysis*.

Data As input we expect a document collection as a set of texts timestamped by publication date. We split the considered time frame into discrete time steps so that each time step is

covered by a substantial set of new documents. While our approach works with any such set of texts, we use a collection of news by CNN on COVID-19-related articles as an application example throughout the whole paper. The texts were scraped based on the script by Pasquali et al., 2021. The dataset consists of 27432 articles, having a median length of 82 words (142 before preprocessing), and a total number of 35544 distinct words (44605 before lemmatization). The articles are all written in English and were published from early 2020 to mid of 2021. Divided into weekly bins, this resulted in 79 time periods, however, with lower coverage towards the end; to avoid time steps without any documents, we considered only the first 76 out of 79 time periods. We have selected this example because everybody will understand the meaning of the concerns discussed, and specific insights can easily be checked with other sources.

Topic Extraction We first conduct common preprocessing steps to format characters to lowercase, remove numbers and punctuation, applied a list of trusted stopwords and a lemmatization dictionary, and removed words with fewer than two characters. We, then, use LDA [Blei et al., 2003b] to derive 12 latent topics in the data. It assigns each occurrence of a word to one of the topics based on its prevalence of occurring together with other words. Words from the same document can be assigned to different topics, since documents may touch upon multiple concerns (Figure 23, left). An overview of the words that occurred in the topics already provides an insight into the topic's main concerns and the total number of words assigned to it reflects its importance, since more important topics are the subject of documents more frequently. In a dynamic collection, however, unmodified LDA is not a good choice as topics would be shuffled with every time step to some extent randomly. Instead, we use a rolling version of LDA [Rieger et al., 2021] to keep topics consistent over time, as words in newly added documents are allocated based on the topics in previous time steps as well. Based on journalism as the domain, we defined a reference period of up to four weeks (or shorter if a change point was detected within that time) weighted its vocabulary with 15% in contrast to the 85% of the new vocabulary of the time step. The parameter selection is rather conservative and leads to comparatively few change points. Figure 23 (right) illustrates, that using this method, we obtain a temporally ordered sequence of ever-changing vocabularies that describe the extracted topics (**AQ 1.1**).

Topic Similarity Although the method maximizes differences between topics, some topics will be more similar than others. Some words might occur so frequently and in such different contexts of other words, that the algorithm assigns them to multiple topics. With topics' concerns being described by their vocabularies, the *pairwise cosine similarity* allows detecting similarities between topics, even if the sizes of the topics are different. We refer to this as *inter-topic similarity*. It can be computed globally across all time steps or locally with respect to a specific time step—the similarity may change as a pair of topics can be similar for a given

time period but deviate again later (**AQ 1.2**). As we are also interested in how the vocabulary of a single topic changes (**AQ 1.1**), the similarity measure can be applied internal to a topic by comparing the vocabulary of the topic to itself at a different point in time (*intra-topic similarity*).

Change Point Detection Even when stabilizing the topics, the vocabulary will never be the same in two consecutive time steps. These changes can be subtle over long periods of time or happen very abruptly from one time step to the next. In order to detect these abrupt changes (**AQ 2.1**), we employ a recently published change point detection method that computes similarities of the topics with respect to a predefined reference period of preceding time steps. It has two parameters that mainly control the sensitivity of the algorithm: the maximum length of the reference period with which a topic at a new time step is compared, and a weighting parameter, which regulates the minimum intensity for a change to be detected.

Change Point Characteristics Based on this predefined reference period and weighting, the cosine similarity of the vocabularies indicates whether the documents from the new time increment have a relevant impact on the topic. A change is detected, if the similarity falls below a dynamically adjusted threshold that accounts for varying topic sizes over time (i.e., with a smaller sample size—the number of words—a topic is expected to vary more). When a topic at a given time step is less similar with its previous states than expected, we can draw multiple pieces of information from it. First, the difference between similarity and threshold indicates how different the vocabularies were in the time step compared to its predecessors. Secondly, the exact values of the similarity and the threshold can contain additional information on whether the topic was stable and we expected a high similarity, or whether it fluctuated, and we already expected a low similarity that was yet undergone. Together with general properties of the topic at this point in time, this provides important characteristics to interpret the change and the circumstances of its detection (**AQ 2.1**).

Change Cause For each time step, and especially for each change point, we can investigate the impact of changed frequencies for certain words, by comparing the similarity of the topic with its previous state when leaving the word out. The words that had the biggest impact, and whether their frequency increased/decreased, and to which degree, can give a thorough impression on why a topic actually changed in the given time step (**AQ 2.2**). Also, shifts of concerns from one topic to another can be detected through such analysis (**AQ 1.3**).

Interlinked Topic Changes We consider the aforementioned aspects to reveal common or related changes within the topics (**AQ 2.3**). For example, a topic that is changed through

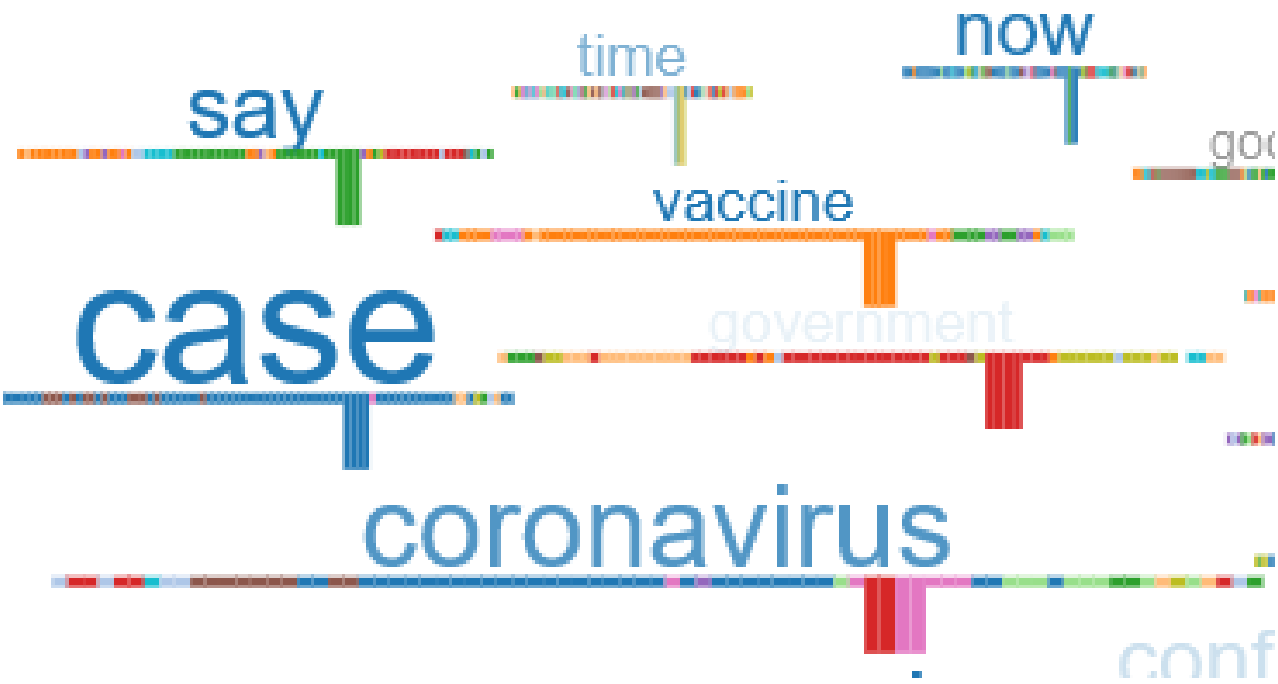


Figure 24: The opacity of words changes based on their relevance to the topic at a selected time step (▮ **AQ 1.1**). While *'coronavirus'* is important in the topic over all, the enlarged segments of the scarfplot show that it was more relevant in other topics at the time. On the other hand, *'case'* is most relevant to the selected topic.

the emergence of a few new words that strongly define the coverage may experience another relevant change if said words disappear from the topic again just as abruptly. Some changes in the documents of a text collection might be large or cross-cutting enough such that they have a relevant impact on multiple topics. Either the emergence of a new word is so present that it starts to occur very frequently in all topics at the same time, or the same new word might impact different topics at different times. Reasoning about change causes may support identifying the words that caused relevant changes not only in different topics, but also at different times.

6.5 Visualization Approach

We propose a visualization approach that enables users to explore the output artifacts of the topic extraction and change detection methods outlined in Section 6.4. The user interface shown in Figure 22 visually links related information and adapts to interactive selection of the color-coded topics, time steps, and words. The approach covers the whole scale from a global overview to the changes in word frequencies at a specific time step in a specific topic.

6.5.1 Adaptive Word Cloud

In the leftmost column of the layout, the word cloud visualization provides an overview of important words in the text collection. The word cloud layout is based on the total number of occurrences of the top 100 most frequent words in the entire dataset. Size encodes the word's frequency on a square root scale with a minimum and maximum value for legibility. The actual width of the word (largely dependent on the number of characters) still influences the area, so that a longer word might still cover more area than a short word, even though its frequency is lower. In addition to size, the frequency—and, hence, importance—is encoded in the position as well, with the words placed greedily towards the top left corner and smaller words filling the gaps between large ones.

Reflecting the temporal changes in which topics a word was used, we introduce a scarfplot below each word. It consists of segments equal to the number of time steps in the dataset. Each segment adopts the color of the topic in which it possesses the highest relevance at the respective time step. Relevance is defined as the share of the word's occurrences relative to the number of total words at the time step (in the topic). This encoding displays whether a word moved from the context of one topic to another over time (▮ **AQ 1.3**).

The word cloud is topic-aware and adjusts the size of the words to the frequency of the words within a selected topic (▮ **AQ 1.1**). Words that surpass a certain frequency threshold within the topic are also colored according to the selected topic to emphasize their relevance. In order to keep a stable mental map, the layout is unchanged. To avoid overlaps between words, this implies that the size of the words cannot grow based on the topic selection and, instead, must be strictly less than before. The relevance of a word in a topic, relative to the overall collection, can be perceived by the degree to which words shrink. The length of the scarfplot always keeps the length of the word's original size to allow for an easier comparison.

If the user selects a time step, the word cloud adapts in two ways as well (Figure 24). First, the opacity of each word changes based on its relevance at the given time step. Words that were used frequently in the overall time span, but rarely at the selected time step, will fade out and leave the words describing the time step more accurately clearly visible (▮ **AQ 1.1**). The scarfplot also highlights the selected time step and its reference period by enlarging the respective segments.

6.5.2 Topic Timelines

At the center of our interface is the list of topic timelines. The topics are sorted by their overall size with the largest topic at the top. Since all topics exist through the whole time span, all

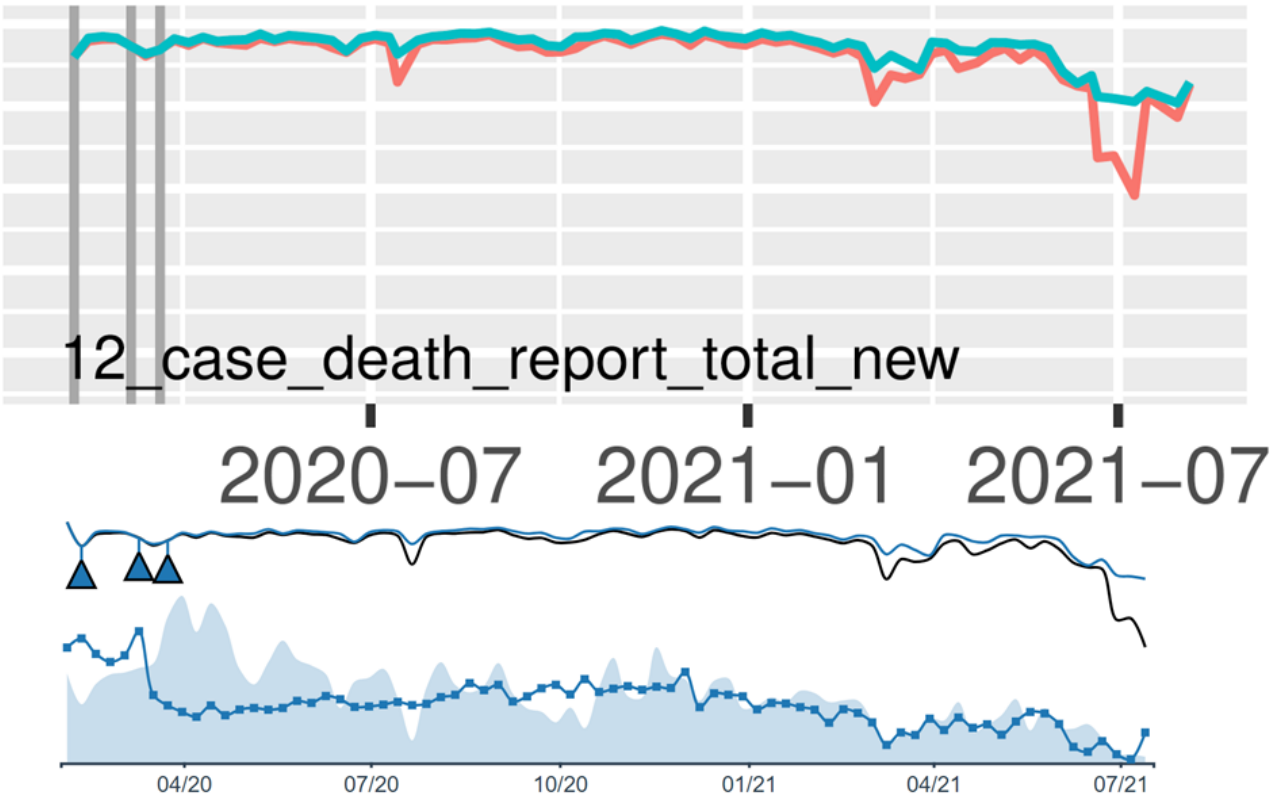


Figure 25: In addition to the similarity and the threshold in the visualization of the original publication (top) Rieger et al., 2022, our timeline visualization adds the size of the topic as well as the relevance of a selected word over time (here: the word *'case'* in topic V12) (**AQ 1.1**).

topics are aligned on the same horizontal axis. Within the timeline we visualize four time series that are also explained in the legend on the left.

The timeline is an extension of the visualizations in the original publication of the statistical methods we use Rieger et al., 2022 (Figure 25). The similarity and dynamic threshold, that were already included in the old version, are calculated based on the reference period in a rolling time window. The similarity is the solid line in the color of the topic, whereas the threshold as the baseline is always black. The detected change points are marked by triangles with leaders to the point at which the similarity fell below the threshold (**AQ 2.1**).

Information we added to the timeline visualization is the size of the topic that represents its importance and is not stable over time (**AQ 1.1**). Depending on the number of documents added per time step and the prevalence of the topic, its size fluctuates. Visually, the topic size is encoded in an area chart in the background. This information is helpful since the value of the threshold is party influenced by the topic size, and this relationship becomes clear from the new visualization.

Moreover, we added functionality to display the relevance of a word in the topic over the course of time. Based on interactive selection in the word cloud (Section 6.5.1) or the change detail

view (Section 6.5.4), the relevance is plotted in a line chart with rectangular glyphs for each time step in which the word occurred in the topic. Comparing the relevance over multiple topic timelines provides a more detailed picture compared to the scarfplot in the word cloud, where each segment is assigned based on a *winner-takes-all* selection (▮ **AQ 2.3**).

Through the timelines, the user can select both, a topic and a time step which is propagated to the word cloud and change detail view as well. If a time step is selected, an indicator will mark it in the timeline, together with its topic-specific reference period of up to four time steps (if no change point was detected within that time).

6.5.3 Temporal Similarity Matrices

On demand, the intra-topic similarity across the whole time span can be accessed (Figure 26, top). On the same horizontal axis as the timeline, the pairwise cosine similarity of the vocabulary between time steps in the same topic is visualized in a matrix. It can be considered the adjacency matrix of a weighted dynamic graph where the edge weights describe the similarity between the time steps as vertices. The intra-topic similarity matrix reveals time spans during which kept a similar vocabulary beyond the reference period and also time spans where vocabulary changed more drastically (▮ **AQ 1.1**).

In order to analyze the similarity between different topics over time, we also introduced the inter-topic similarity matrix (Figure 26, bottom). Every topic has a row and a column. Cells where different topics intersect in the top half, contain the 95-percentile of similarities between the two topics. Using the percentile instead of the average, points out strong relationships more clearly. However, the exact choice of the percentile depends on the dataset (e.g., the number of time steps). The bottom half of the matrix shows the course of similarities between the two topics over time and reveals time spans where two topics were particularly similar or dissimilar (▮ **AQ 1.2**).

6.5.4 Change Detail View

When the user selects a topic and a time step, the most relevant changes in word frequencies are shown in detail, even if no change point was detected. Since change points are triggered by gaping cosine similarities, seeing how individual words impacted this gap is helpful in understanding how the topic changed. The 30 words with the highest impact on the similarity change—based on *leave-one-out* calculations—represent the change cause and are listed in the top right of the system (▮ **AQ 2.2**).

For each word in the list, we contrast the frequency of the word in the selected time step with the average frequency in the reference period. Based on whether the frequency increased or

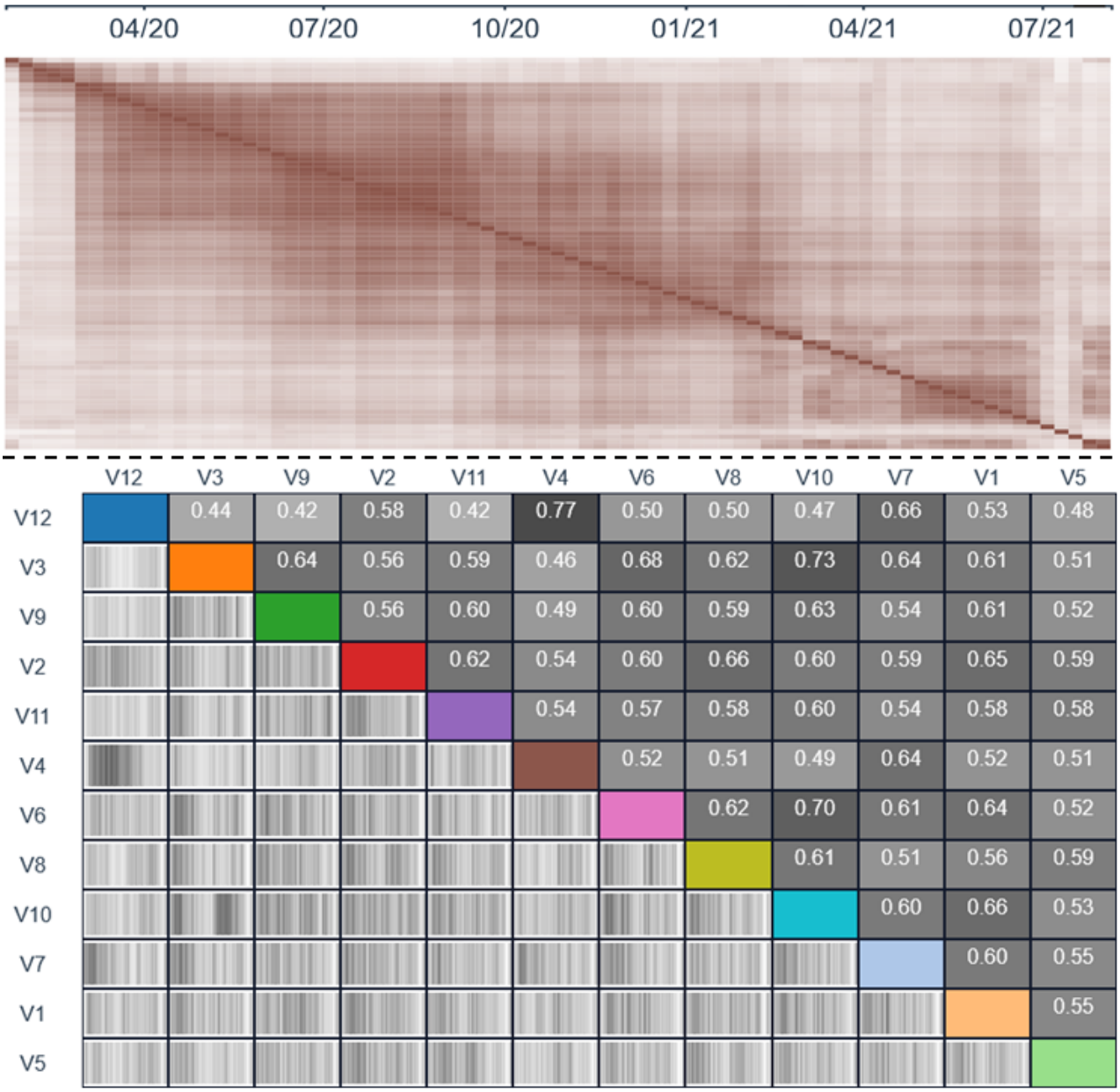


Figure 26: The intra-topic similarity matrix (top) shows the pairwise similarity between time steps within the same topic (**AQ 1.1**). The inter-topic similarity matrix (bottom) shows the pairwise similarity between topics at the same time step over time in the bottom half and the 0.95-percentile value in the top half (**AQ 1.2**).

decreased, a bar is placed on the right or left of a butterfly chart with the length of the bar showing the word's impact on the cosine similarity in the time step (Figure 27).

6.6 Change Patterns

During the use of the system and as part of the development process, we have identified a set of specific patterns within the topics. The identification typically requires looking at the

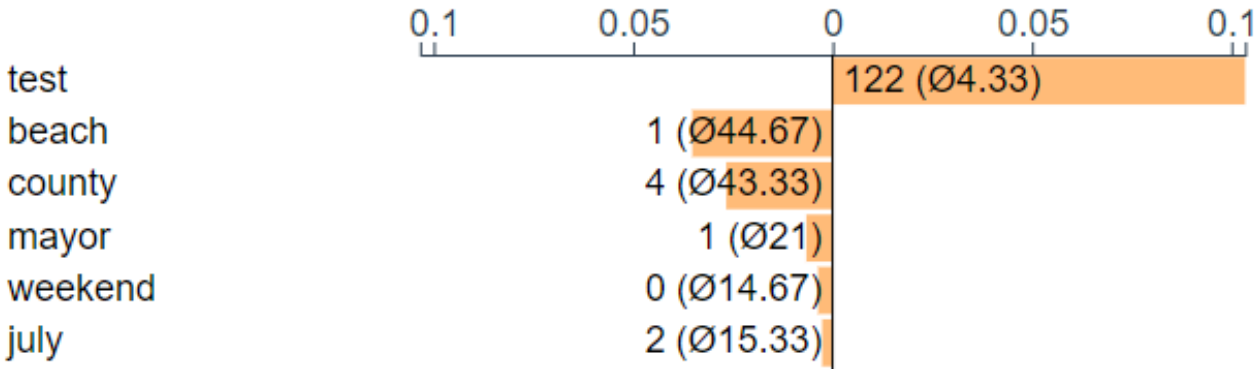


Figure 27: The words with the highest impact on similarity that led to the detection of the change point in time step 25 in topic V1 (**AQ 2.2**).

data from multiple perspectives in different visualizations and combining answers of several analysis questions. In the following, we describe the patterns along their data characteristics and separate them by the number of detected change points in the pattern—from none, over single change points, to multiple ones. Some patterns might occur in conjunction as well. Individual change points within a multi-change point pattern might follow a single change point pattern if the observation interval is slightly adjusted.

No Change Point Whenever we want to detect substantial changes, by definition, there are states in which the topics are rather stable. Our approach enables users to also explore the topics at time steps where no change point was detected. We were able to identify three distinct behaviors exhibited in the topics.

- **Stable behavior:** When no abrupt change points are detected within a certain time window, it is straightforward to assume that a topic stayed stable. It is typically indicated by both, high absolute values for similarity and threshold, as well as few dips in these time series. It can additionally be confirmed by investigating the intra-topic similarity.
- **Volatile behavior:** Some topics derived by the model do not have sharp boundaries in certain time windows. They can either be inherently volatile (e.g., political coverage concerning daily events) that have a few constant elements (e.g., political actors), or they can be a loose composition of words that did not fit well into the other topics. Topics with such a behavior commonly display comparatively low absolute values for similarity and threshold and frequent swings in them. This pattern is also sensitive to the selection of model parameters.
- **Slow topic drift:** The dynamic change detection method we use is specifically designed to identify abrupt changes. Hence, it is possible that topics change over time, but the manifestation of that trend—in each time step alone—is small enough not to trigger a

change point. The frequencies of important words in the topic at a time step slowly decrease over time, while the frequencies of new words slowly increase. This change often occurs over the course of larger time spans, and it is quite sensitive to the parameter selection (e.g., the length of the reference period). For drifting topics, it is further interesting to investigate whether other topics take up parts of the drifting aspects.

Single Change Point A change point is detected when the manifested similarity of a topic at a time step with its reference period of previous states is lower than it was expected considering the stability and size of the topic in those periods. Such a difference is caused by substantial changes in the topic's concerns. The impact of each individual word at the change point can reveal more details about what exactly caused the change.

- **Single concern dominance:** Some change points are caused by the changed frequency of a single word, indicating that the appearance or disappearance of its concern were very unexpected in the topic at that point in time. Typically, other words also changed in frequency, but with a much smaller impact on the cosine similarity. Closely related, some change points might be caused by multiple words that have much larger impacts than all other words, but they belong to the same concern (e.g., the words *'mask'* and *'mandate'*).
- **Multi-concern change:** In contrast to single topic dominance, change points can also be caused by the impact of changed frequencies of multiple words. These words seem to originate from different concerns that appear or disappear together, or, alternatively, a subset of the words decreases in frequency while the remaining words increase in frequency.

Multiple Change Points Sometimes multiple change points are connected syntactically or semantically. This information is lost when investigating each change point only individually. In particular, we discern three expressions of relationships between multiple change points.

- **Toggle:** Sometimes the coverage of a new concern that constitutes a change point is so extensive that the words describing the concern become defining for the topic at the time step. If said concern disappears later (either quickly or after multiple time steps) it causes another change point with a similar set of words as the initial change point. Oftentimes, the new concern enters at a time step when the topic grows sharply.
- **Multi-step progression:** Some concerns evolve over time. Their importance to the topic is so large that the changes in the words' frequencies are sufficient to trigger multiple change points. This pattern is characterized by the fact that successive change points describe a similar development (e.g., changes are triggered in consecutive time steps that concern first the closing of restaurants, then the closing of schools, and then the ban

Table 24: Mapping of topic change patterns to visualization views and analysis questions; relevance: • *clearly relevant*, ○ *partly relevant*.

Change pattern	Cloud	Timeline	Matrices	Detail	AQ 1.1	AQ 1.2	AQ 1.3	AQ 2.1	AQ 2.2	AQ 2.3
No Change Point										
Stable behavior		●	○		●					
Volatile behavior		●			●					
Slow topic drift	○	○	●		●	○	○			
Single Change Point										
Single concern dominance		○		●	○				●	
Multi-concern change		○		●	○				●	
Multiple Change Points										
Toggle		●	●	●	○	○		●	●	●
Multi-step progression		●		●	○			●		○
Synchronous change	○	●	○	○	○		●	●	○	●

of airplane travel). The major words are, typically, different for all change points and, therefore, the detection of this pattern requires domain knowledge.

- **Synchronous change:** Several instances of the same words can be used in different contexts and are then likely assigned to multiple topics in the same time step. If the appearance of these words is abrupt in several topics, it is possible for change points in different topics to be traced back to the same source. Another manifestation of this pattern is different words causing change points in different topics, although they stem from the same concern.

Table 24 provides details on the specific mapping of change patterns to the visualizations and analysis questions that are relevant for investigating the respective pattern. Consideration of multiple visualizations is, typically, required to identify the change patterns. The fact that the majority of patterns maps to multiple analysis questions emphasis their high-level nature. The timeline visualization contains the change point markers and, since we define the change patterns on the basis of change points in them, is naturally relevant across the board. The change detail view as the source of information on explaining change points is just as straightforwardly tied to the change patterns that contain at least one change point. On the other hand, the word cloud visualization and analysis question **AQ 1.2** are not represented strongly by the change patterns. The word cloud displays an overview of the whole time span or can focus on a snapshot of the concerns at a particular time step, but its only capabilities of representing change itself is in the scarfplots. Similarly, **AQ 1.2** is not targeted at specific temporal changes and is more cross-cutting.

6.7 Results

We analyze the data, discussed in Section 6.4, across the multiple linked visualizations in order to find noteworthy phenomena. To this end, we use the analysis questions from Section 6.3 and the change patterns from Section 6.6. While the findings are based on real world concerns in the dataset, we point out how our approach enabled us to identify these behaviors that were previously hidden.

First, we can get an overview of the major concerns in the text collection by looking at the word cloud visualization in its initial state. With the biggest and most frequent words in the collection towards the top left, we quickly observe many COVID-19-related words like *'coronavirus'*, *'case'*, and *'vaccine'*. Looking further, we also identify words that give more context, like *'us'* (United States), or *'test'*. Considering all words, we can infer that the text collection comprises different perspectives, be it geographical or others.

Looking at the area chart that encodes the topic size in each timeline, we see that there are no major differences in overall topic size, but they can differ substantially at specific time steps. We can also see the overall trend of topics being larger in the beginning and shrinking with time—albeit with some noise in every topic—, which was not visible with the visualization techniques used in the original publication of the change detection method, as we can see in Figure 25 (▮ **AQ 1.1**). The biggest topic (V12) at the top of our list is very stable with respect to similarity and its threshold. It has three change points at the beginning, all of which with a minimal difference between similarity and threshold. Afterwards, it displays the **stable behavior** change pattern, with high absolute similarity values, no observable trend across time in terms of most impactful words (▮ **AQ 1.1**). The intra-topic similarity matrix for V12 shows a very uniform high similarity throughout the time span. With the topic selected, the topic-adjusted word cloud, supports this rationale. Major words are *'case'*, *'death'*, and *'report'*, which we would expect to find in a plain, rational topic that mostly concerns the reporting of numbers. The reoccurring reporting articles have become normal during the pandemic and, hence, observing this as a stable topic aligns with our expectations.

In the inter-topic similarity matrix, we observed that the reporting topic V12 had a high similarity with topic V4 in the first half of the overall time span (▮ **AQ 1.2**). The adjusted word cloud for topic V4 shows *'case'* as an important word. Through the scarfplot, we can infer that, while the word was most relevant in V12, V4 utilized the word even more during a few periods of the first half. Selecting the word, we see that its importance is high in the first half and slowly decreases from around September 2020. The topic's other important word *'state'*, however, stays important through the whole topic. Interactively investigating more time steps and their highest similarity impacts in the change detail view, we also see frequently changing US state names (Figure 28). We inferred that topic V4 mostly concerns US news on

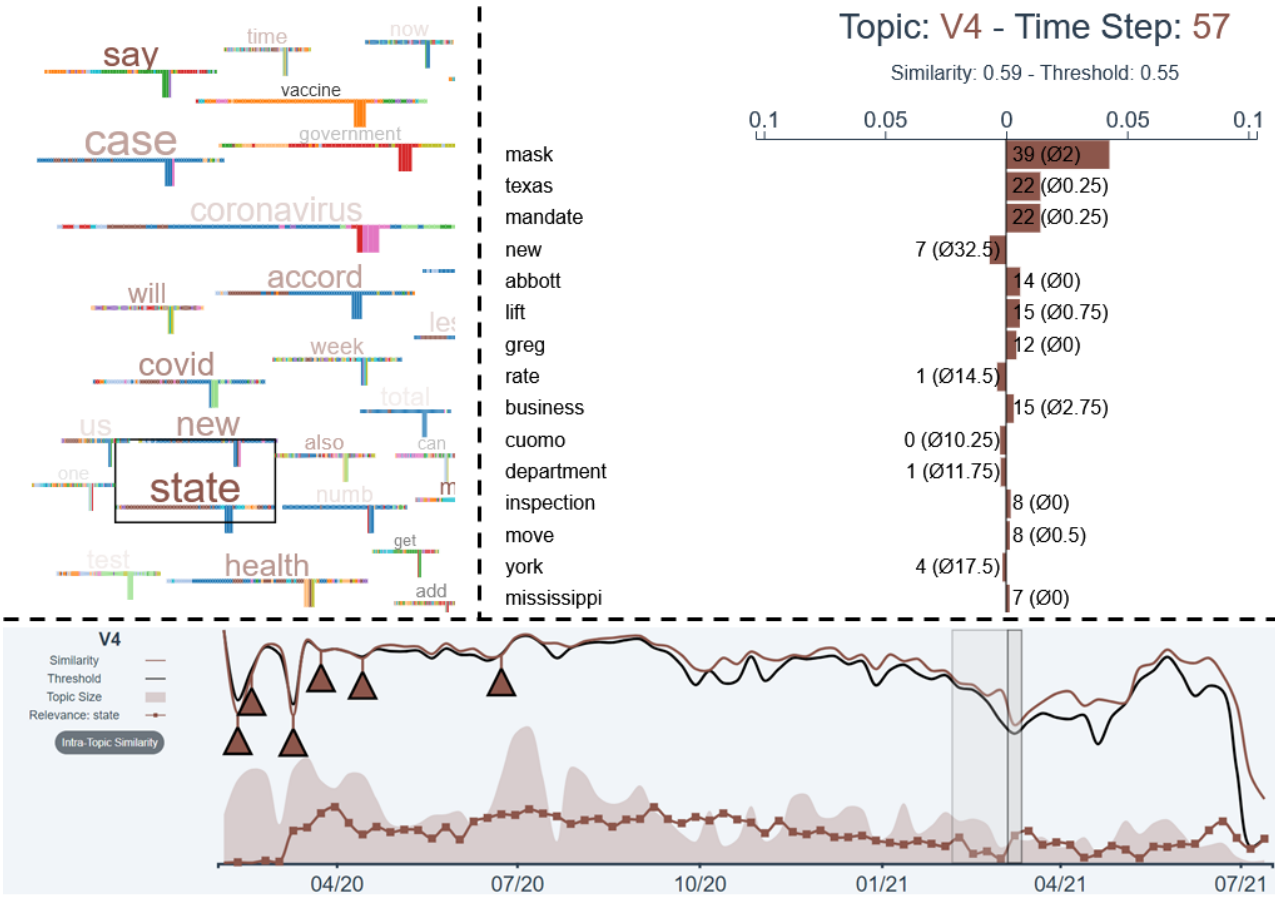


Figure 28: While *'case'* is relevant in topic V4 considering the whole time span, it becomes less important towards the end. The word *'state'* remains important throughout.

state-level and exemplifies a **slow topic drift** pattern from reporting in the beginning (which constituted the similarity to V12) to more general news with more time passed (▮ **AQ 1.1**). This behavior can also be observed in the intra-topic similarity matrix in Figure 26 (top), which shows ever decreasing similarity the higher the distance between two time steps, even though no change points were triggered for a long time span.

Two topics that have different characteristics from the aforementioned ones are V1 and V9. In general, we can see many more change points in the topic timelines (▮ **AQ 2.1**). Investigating major words in the topic-adjusted word clouds, we found them to be political topics from the UK (V1) and US (V9) (▮ **AQ 1.1**). The fast-paced coverage of political news causes these topics to have a lower overall intra-topic similarity, which is visible in the similarity matrices within the topics as well as the similarity and threshold time series. Figure 29 shows that even in times of no detected change points, these topics display the pattern of **volatile behavior**.

Topic V8 concerns many stories about events in different countries and their implications on travel. The nature of such events to occur abruptly is reflected in the change points as well. They are distributed quite evenly over the time span, with relatively stable periods in

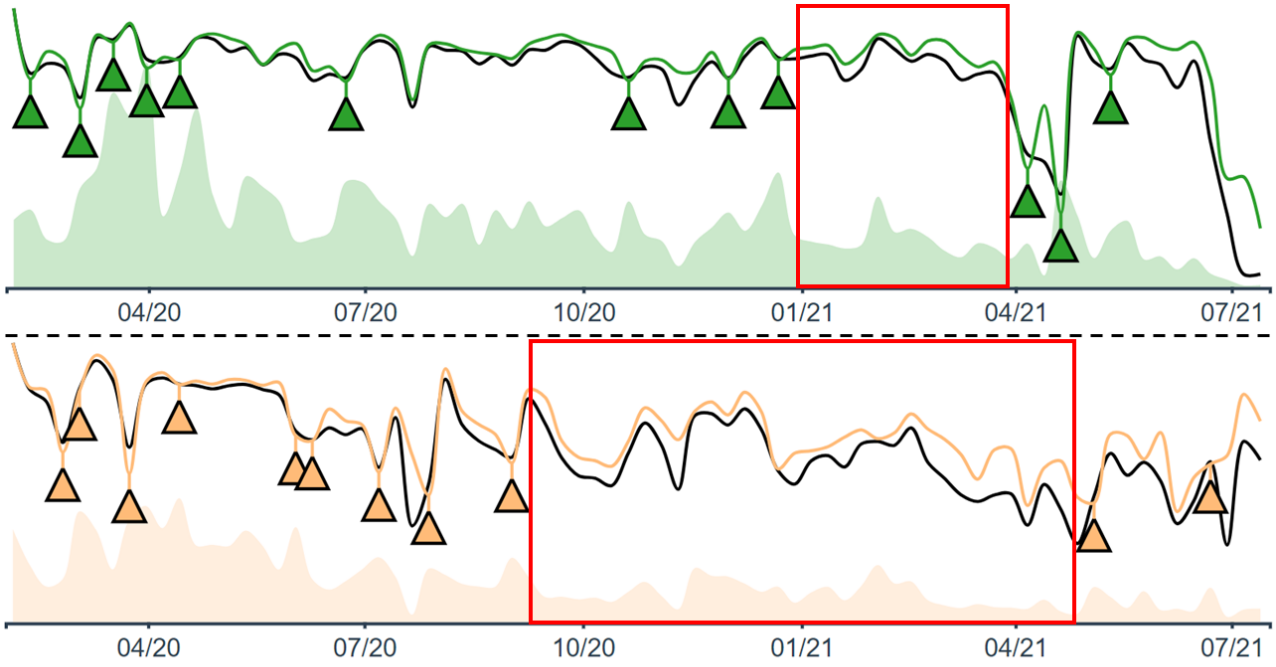


Figure 29: The similarity and threshold line for topics V9 (top) and V1 (bottom) are fluctuating but no change point is detected, which is an indicator for the **volatile behavior** pattern.

between (▮ **AQ 2.1**). The change points themselves are then mostly following the **single story dominance** pattern (▮ **AQ 2.3**), which we can clearly see in the butterfly chart of the change detail view. Examples are the sudden shortage of oxygen supplies in Indian hospitals in early May 2021, or the identification of a new coronavirus UK variant in December 2020 (▮ **AQ 2.2**), both in topic V8. Selecting the word *'variant'*, we observed its rise in topic V8, but also in topic V6. Similarly, we identified another time step in the topic timelines at which a single story impacted multiple topics in December 2020, when the first vaccines were available and the vaccination process was rolled-out. This story constituted **synchronous changes** big enough to trigger change points at the same time step in V9 and V10, as we confirmed in the change detail view (▮ **AQ 2.1**, ▮ **AQ 2.3**).

During the analysis, we noticed a large amount of **toggle** change patterns. Words leaving the topic in one change point are also the major cause (▮ **AQ 2.3**) of an earlier change point within its reference period, as we can observe with the time step indicator in the topic timeline (▮ **AQ 2.1**). We identified cases where the second change point occurred just one time step after the first, like in the case of UK politician Dominic Cummings violating restrictions with a trip to Durham in May 2020 within topic V1. A week later, the absence of the story caused the next change point (▮ **AQ 2.2**). In other cases, there were 2-4 weeks before the second change point of the **toggle** pattern, like in the case of the words *'bill'* and *'stimulus'* in March/April 2020 in V9 (Figure 30) or *'beach'* and *'county'* in July 2020 in V1. The second change point of the latter case also exemplifies the **multi-story change** with the word *'test'* seemingly replacing

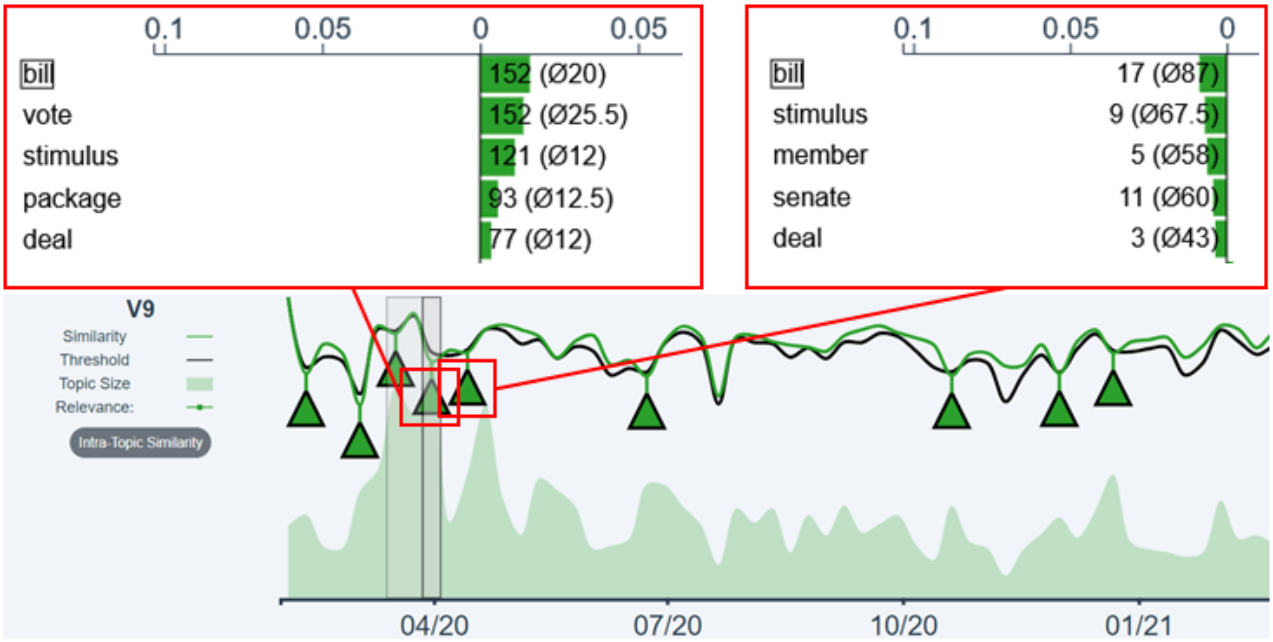


Figure 30: Detailed inspection reveals the **toggle** pattern, as the same words cause change points with their appearance and disappearance in quick succession in the same topic. Both change points individually also represent the **singe-story dominance** pattern.

'beach' and *'county'*, which disappeared, as is visible in the change detail view (**AQ 1.1**, **AQ 2.2**) (Figure 27). Within all **multi-story changes**, we only identified those where one story left the topic and another entered. We attribute it to LDA that not once two new topics enter the same topic at the same time step.

The multiple change points (**AQ 2.1**) at the beginning in topic V11 are related to each other as they describe developments on restrictions like closings of schools and restaurants (**AQ 2.2**) in a **multi-step progression** of the topic. The stories belong to a real-world concern that is not visible in the topic extraction. Hence, the user needs to bring domain knowledge for the confirmation of that pattern, because the shared characteristics are only implicit (**AQ 2.3**). A multi-step progression of the same words rising consecutively is not possible in our data, since the selection of cosine similarity makes high-relevance words triggering change points by becoming even more frequent unlikely.

6.8 Discussion and Research Opportunities

Since our contributions are the development and demonstration of a visualization approach, we will abstract our discussion from limitations of the used topic extraction and change detection algorithms. We focus more on the implications of our approach supporting topic extraction

experts to understand their methods and how other users could benefit from our approach as well.

Analysis Questions and Change Patterns Through constant exchange and design iterations, we were able to create a visualization approach that is capable of answering the analysis questions we identified in Section 6.3. Within the presentation of our results, we showed that our analysis questions unveiled deep insights into the output artifacts and helped the experts understand the implications of certain parameter choices. The only exception is **AQ 1.3**, which we could not observe in our analysis. Our statistics experts attributed this to the intricacies of LDA as a topic extraction technique. We believe if we have moving concerns between topics, that it would be visible in our visualization system as a **synchronous change** pattern. Our focus on the analysis of detected change points (**AQ 2**), in general, revealed rich opportunities for analysis and even led us to the categorization of method-agnostic change patterns we observed while using the system. The fact that these patterns are latent in the data, but the algorithm is not designed to explicitly detect these, informs the need for further developments in this direction. Once a method for the automated identification of these patterns is available, its results could be well integrated into our approach. In that sense, the current point-in-time interpretation of change points could be extended to also allow changes along a time interval.

Generalizability While our implementation is somewhat tied to the specific outputs of an LDA topic extraction model and the detection of abrupt change points, we believe many of our contributions are more broadly applicable to other methods as well. The high-level analysis questions and change patterns we have identified are independent of the used method and only rely on the detection of change points in general. Furthermore, the comparative word clouds, and timeline visualizations can likely be ported to other methods right away. The temporal correlation matrices within and between topics, as well as the detail inspection of detected changes could be adjusted to display the same or similar information for other models with a few tweaks. The matrices could allow a changing number of topics over time by employing dynamic graph visualization techniques [Beck et al., 2016]. The list of the most impacting words at a given time step can be extended to intervals of time to account for approaches that detect gradual changes as opposed to abrupt ones.

Insights for End Users Change point detection is typically done online (i.e., on data streaming at the time of the analysis) in order to identify insights about the data. We, however, do not assume a perfect method and want to enable experts to tweak it according to insights collected through retrospective analysis of the detection method. We designed the visualization approach together with researchers of topic extraction and change point detection methods to

assist them in improving the understanding and calibration of their methods depending on the research question in the application under study. Nonetheless, the results we presented in Section 6.7, already indicate that actual insights about the underlying data can be gathered as well. We believe that our approach can be adapted to assist users interested in a text collection itself (e.g., literary scholars) in gathering insights about the contents of the collection. Instead of looking for improvements in the topic extraction and change detection, we would then assume a sufficient quality of the outputs to support these non-technical users. This would imply to reduce complexity from the method comprehension side of the approach and focus on domain-specific aspects like the inclusion of multiple text sources to compare the topics across those boundaries. Another important step in this case would be to integrate the original text documents into the approach itself. We have discussed this option, but the integration of the documents poses many new challenges, like finding representative documents from the large corpus to assign to the topics at any given time step. Moreover, the fact that, with LDA, not the documents themselves are assigned to the topics, but the words within the documents, makes it even more difficult to make such a selection.

6.9 Conclusion

In this paper we proposed a visualization approach to support the comprehension of change points in temporal text collections. We defined the abstract problem space along analysis questions in order to help examine the role of change points in dynamic topic extraction, and discussed our concrete application scenario in the context of the questions. We developed a visualization approach to analyze and connect output artifacts of the analysis methods, and demonstrated the usefulness of the approach through an in-depth analysis of a temporal text collection on COVID-19-related news data. Through our analysis we were able to identify specific instances that could inform parameter choices for the method's calibration, based on desired properties in the application scenario. Furthermore, we generated insights on the derived topics, as well as their contents over time. From our observations, we classified eight high-level change patterns from stable time periods to multi-step progressions spanning across detected change points. We specifically see opportunities for future research in the automated detection of such high-level patterns and their more explicit visual presentation to a wider base of users.

7

Conclusion

This cumulative thesis investigated and developed statistical methods for complex data structures. The main focus was on network data, which is becoming increasingly important with the rise of big data and artificial intelligence. Chapters 2 and 3 developed an online change detection procedure that can be used in flexible circumstances without making high-level assumptions about the network structure. Although the loss of information is mitigated by the multivariate metric-based approach introduced in Chapter 3, a topic for future research may be to find a way to make also model-based approaches usable in more flexible situations by relaxing their restrictions. This would include allowing node dynamics - not only at low intensities, for which there exist already some straightforward strategies such as expanding the adjacency matrix with 0-rows or offset terms [Krivitsky et al., 2011], but also at higher intensities. Regarding the dynamic network setup for online monitoring, another issue is the time scale, as snapshots of the network are taken at particular time points, resulting in a somewhat discrete consideration of graphs. However, this can only be seen as an approximation of the dynamics of real networks, which typically evolve continuously over time. The general framework provided in Nguyen et al., 2018 can be used as a basis for more sophisticated analysis in this regard.

Further opportunities for future research include the derivation of asymptotics for certain network metrics such as centrality scores. Their conception as graph functionals as in Chapter 4 is only a first step in this direction. Related to this topic is the derivation of closed forms of moments. Due to the complex nature of these metrics, it is doubtful that such a closed form can be derived, but even an approximation would be of great value. For now, resampling methods such as the bootstrap tests in Chapter 4 offer an alternative.

Lastly, since network data is basically a result of the growing technological progress and the associated richer information content of data, a major focus of research should be on scalability of novel and existing methods. That is, making statistical approaches effective and practical not only for small and medium-sized networks, but also for very large networks that are the rule rather than the exception in modern times.

Bibliography

Adams, Ryan Prescott and David JC MacKay (2007). *Bayesian online changepoint detection*. arXiv: 0710.3742.

Banerjee, Sayantan and Kousik Guhathakurta (2020). "Change-point analysis in financial networks". In: *Stat* 9.1, e269. DOI: 10.1002/sta4.269.

Barnett, Ian and Jukka-Pekka Onnela (2016). "Change point detection in correlation networks". In: *Scientific reports* 6. DOI: 10.1038/srep18893.

Bassett, Danielle S and Olaf Sporns (2017). "Network neuroscience". In: *Nature neuroscience* 20.3, pp. 353–364. DOI: 10.1038/nn.4502.

Basseville, Michèle and Igor V Nikiforov (1993). *Detection of abrupt changes: theory and application*. Vol. 104. Prentice Hall Englewood Cliffs.

Beck, Fabian, Michael Burch, Stephan Diehl, and Daniel Weiskopf (2016). "A Taxonomy and Survey of Dynamic Graph Visualization". In: *Computer Graphics Forum* 36.1, pp. 133–159. DOI: 10.1111/cgf.12791.

Belkin, Mikhail and Partha Niyogi (2001). "Laplacian eigenmaps and spectral techniques for embedding and clustering". In: *Advances in neural information processing systems* 14.

Bhattacharyya, Sharmodeep and Peter J Bickel (2015). "Subsampling bootstrap of count features of networks". In: *The Annals of Statistics* 43.6, pp. 2384–2411. DOI: 10.1214/15-AOS1338.

Bickel, Peter J and Purnamrita Sarkar (2016). "Hypothesis testing for automated community detection in networks". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.1, pp. 253–273. DOI: 10.1111/rssb.12117.

Blei, David M., Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum (2003a). "Hierarchical topic models and the nested Chinese restaurant process". In: *Advances in Neural Information Processing Systems*. Vol. 16. MIT Press, pp. 17–24. URL: https://proceedings.neurips.cc/paper/2003/hash/7b41bfa5085806dfa24b8c9de0ce567f-Abstract.html.

Blei, David M. and John D. Lafferty (2006). "Dynamic Topic Models". In: *Proceedings of the 23rd ICML-Conference*. ACM, pp. 113–120. DOI: 10.1145/1143844.1143859.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003b). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022. DOI: 10.1162/jmlr.2003.3.4-5.993.

Bose, Avinandan and Soumendu Sundar Mukherjee (2021). *Changepoint Analysis of Topic Proportions in Temporal Text Data*. arXiv: 2112.00827.

Brouwer, Andries E and Willem H Haemers (2011). *Spectra of graphs*. Springer Science & Business Media.

Bubeck, Sébastien, Jian Ding, Ronen Eldan, and Miklós Z Rácz (2016). "Testing for high-dimensional geometry in random graphs". In: *Random Structures & Algorithms* 49.3, pp. 503–532. DOI: 10.1002/rsa.20633.

Bunke, Horst, Peter J Dickinson, Miro Kraetzl, and Walter D Wallis (2007). *A graph-theoretic approach to enterprise network dynamics*. Vol. 24. Springer Science & Business Media.

Burch, Michael, Steffen Lohmann, Fabian Beck, Nils Rodriguez, Lorenzo Di Silvestro, and Daniel Weiskopf (2014). "RadCloud: Visualizing multiple texts with merged word clouds". In: *Proceedings of the International Conference on Information Visualisation*, pp. 108–113. DOI: 10.1109/IV.2014.72.

Carrington, Peter J, John Scott, and Stanley Wasserman (2005). *Models and methods in social network analysis*. Vol. 28. Cambridge university press.

Chapanond, Anurat, Mukkai S Krishnamoorthy, and Bülent Yener (2005). "Graph theoretic and spectral analysis of Enron email data". In: *Computational & Mathematical Organization Theory* 11.3, pp. 265–281. DOI: 10.1007/s10588-005-5381-4.

Chen, Li, Jie Zhou, and Lizhen Lin (2021). "Hypothesis testing for populations of networks". In: *Communications in Statistics-Theory and Methods*, pp. 1–24. DOI: 10.1080/03610926.2021.1977961.

Cheung, Rex CY, Alexander Aue, Seungyong Hwang, and Thomas CM Lee (2020). "Simultaneous Detection of Multiple Change Points and Community Structures in Time Series of Networks". In: *IEEE Transactions on Signal and Information Processing over Networks* 6, pp. 580–591. DOI: 10.1109/TSIPN.2020.3012286.

Coe, Neil M, Peter Dicken, and Martin Hess (2008). "Global production networks: realizing the potential". In: *Journal of economic geography* 8.3, pp. 271–295. DOI: 10.1093/jeg/lbn002.

Cui, Peng, Xiao Wang, Jian Pei, and Wenwu Zhu (2018). "A survey on network embedding". In: *IEEE transactions on knowledge and data engineering* 31.5, pp. 833–852. DOI: 10.1109/TKDE.2018.2849727.

Cui, Weiwei, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong (2011). "TextFlow: Towards Better Understanding of Evolving Topics in Text". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12, pp. 2412–2421. DOI: 10.1109/tvcg.2011.239.

Cui, Weiwei, Shixia Liu, Zhuofeng Wu, and Hao Wei (2014). "How Hierarchical Topics Evolve in Large Text Corpora". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12, pp. 2281–2290. DOI: 10.1109/tvcg.2014.2346433.

Dan, Soham and Bhaswar B Bhattacharya (2020). "Goodness-of-fit tests for inhomogeneous random graphs". In: *International Conference on Machine Learning*. PMLR, pp. 2335–2344.

Dou, Wenwen and Shixia Liu (2016). "Topic- and Time-Oriented Visual Text Analysis". In: *IEEE Computer Graphics and Applications* 36.4, pp. 8–13. DOI: 10.1109/mcg.2016.73.

Dou, Wenwen, Xiaoyu Wang, Drew Skau, William Ribarsky, and Michelle X. Zhou (2012). "LeadLine: Interactive Visual Analysis of Text Data through Event Identification and Exploration". In: *IEEE Conference on Visual Analytics Science and Technology (VAST)*. DOI: 10.1109/vast.2012.6400485.

Duan, Dongsheng, Lingling Tong, Yangxi Li, Jie Lu, Lei Shi, and Cheng Zhang (2020). "Aane: Anomaly aware network embedding for anomalous link detection". In: *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, pp. 1002–1007. DOI: 10.1109/ICDM50108.2020.00116.

Durante, Daniele and David B Dunson (2014). "Bayesian dynamic financial networks with time-varying predictors". In: *Statistics & Probability Letters* 93, pp. 19–26. DOI: 10.1016/j.spl.2014.06.015.

Durante, Daniele, David B Dunson, and Joshua T Vogelstein (2017). "Nonparametric Bayes modeling of populations of networks". In: *Journal of the American Statistical Association* 112.520, pp. 1516–1530. DOI: 10.1080/01621459.2016.1219260.

Efron, Bradley (1992). "Bootstrap methods: another look at the jackknife". In: *Breakthroughs in Statistics: Methodology and Distribution*. Springer, pp. 569–593.

Farahani, Ebrahim Mazrae, Reza Baradaran Kazemzadeh, Rassoul Noorossana, and Ghazaleh Rahimian (2017). "A statistical approach to social network monitoring". In: *Communications in Statistics-Theory and Methods* 46.22, pp. 11272–11288. DOI: 10.1080/03610926.2016.1263741.

Flossdorf, Jonathan, Roland Fried, and Carsten Jentsch (2023). "Online monitoring of dynamic networks using flexible multivariate control charts". In: *Social Network Analysis and Mining* 13.1, p. 87. DOI: 10.1007/s13278-023-01091-y.

Flossdorf, Jonathan and Carsten Jentsch (2021). "Change Detection in Dynamic Networks Using Network Characteristics". In: *IEEE Transactions on Signal and Information Processing over Networks* 7, pp. 451–464. DOI: 10.1109/TSIPN.2021.3094900.

Forster, Peter, Lucy Forster, Colin Renfrew, and Michael Forster (2020). "Phylogenetic network analysis of SARS-CoV-2 genomes". In: *Proceedings of the National Academy of Sciences* 117.17, pp. 9241–9243. DOI: https://doi.org/10.1073/pnas.2004999117.

Frermann, Lea and Mirella Lapata (2016). "A Bayesian Model of Diachronic Meaning Change". In: *Transactions of the Association of Computational Linguistics* 4, pp. 31–45. DOI: 10.1162/tacl_a_00081.

Fryzlewicz, Piotr (2014). "Wild binary segmentation for multiple change-point detection". In: *The Annals of Statistics* 42.6, pp. 2243–2281. DOI: 10.1214/14-AOS1245.

Gad, Samah, Waqas Javed, Sohaib Ghani, Niklas Elmqvist, Tom Ewing, Keith N. Hampton, and Naren Ramakrishnan (2015). "ThemeDelta: Dynamic Segmentations over Temporal Topic Models". In: *IEEE Transactions on Visualization and Computer Graphics* 21.5, pp. 672–685. DOI: 10.1109/tvcg.2014.2388208.

Gao, Chao and John Lafferty (2017). "Testing network structure using relations between small subgraph probabilities". In: *ArXiv preprint arXiv:1704.06742*.

Gao, Chao, Yu Lu, and Harrison H. Zhou (2015). "Rate-optimal graphon estimation". In: *The Annals of Statistics* 43.6, pp. 2624–2652. DOI: 10.1214/15-AOS1354. URL: https://doi.org/10.1214/15-AOS1354.

Ghoshdastidar, Debarghya and Ulrike von Luxburg (2018). "Practical methods for graph two-sample testing". In: *Advances in Neural Information Processing Systems*, pp. 3019–3028.

Gilbert, Edgar N (1959). "Random graphs". In: *The Annals of Mathematical Statistics* 30.4, pp. 1141–1144.

Gobbo, Beatrice, Duilio Balsamo, Michele Mauri, Paolo Bajardi, André Panisson, and Paolo Ciuccarelli (2019). "Topic Tomographies (TopTom): a Visual Approach to Distill Information from Media Streams". In: *Computer Graphics Forum* 38.3, pp. 609–621. DOI: 10.1111/cgf.13714.

Grattarola, Daniele, Daniele Zambon, Lorenzo Livi, and Cesare Alippi (2019). "Change detection in graph streams by learning graph embeddings on constant-curvature manifolds". In: *IEEE Transactions on neural networks and learning systems* 31.6, pp. 1856–1869. DOI: 10.1109/TNNLS.2019.2927301.

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl 1, pp. 5228–5235. DOI: 10.1073/pnas.0307752101.

Grover, Aditya and Jure Leskovec (2016). "node2vec: Scalable feature learning for networks". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864. DOI: 10.1145/2939672.2939754.

Gürsoy, Furkan, Mounir Haddad, and Cécile Bothorel (2021). "Alignment and stability of embeddings: measurement and inference improvement". In: *arXiv preprint arXiv:2101.07251*.

Haddad, Mounir, Cécile Bothorel, Philippe Lenca, and Dominique Bedart (2020). "Temporalnode2vec: Temporal node embedding in temporal networks". In: *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*. Springer, pp. 891–902. DOI: 10.1145/2939672.2939751.

Havre, Susan, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell (2002). "ThemeRiver: Visualizing Thematic Changes in Large Document Collections". In: *IEEE Transactions on Visualization and Computer Graphics* 8.1, pp. 9–20. DOI: 10.1109/2945.981848.

Hazrati-Marangaloo, Hossein and Rassoul Noorossana (2021). "A nonparametric change detection approach in social networks". In: *Quality and Reliability Engineering International* 37.6, pp. 2916–2935. DOI: 10.1002/qre.2897.

Hewapathirana, Isuru U (2019). "Change detection in dynamic attributed networks". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.3, pp. 1286–1306. DOI: 10.1002/widm.1286.

Hewapathirana, Isuru Udayangani, Dominic Lee, Elena Moltchanova, and Jeanette McLeod (2020). "Change detection in noisy dynamic networks: a spectral embedding approach". In: *Social Network Analysis and Mining* 10, pp. 1–22. DOI: 10.1007/s13278-020-0625-3.

Hoeffding, Wassily (1992). "A class of statistics with asymptotically normal distribution". In: *Breakthroughs in Statistics*. Springer, pp. 308–334.

Holland, Paul W, Kathryn Blackmond Laskey, and Samuel Leinhardt (1983). "Stochastic blockmodels: First steps". In: *Social networks* 5.2, pp. 109–137. DOI: 10.1016/0378-8733(83)90021-7.

Hu, Jianwei, Jingfei Zhang, Hong Qin, Ting Yan, and Ji Zhu (2021). "Using maximum entry-wise deviation to test the goodness of fit for stochastic block models". In: *Journal of the American Statistical Association* 116.535, pp. 1373–1382. DOI: 10.1080/01621459.2020.1722676.

Huberman, Bernardo A, Daniel M Romero, and Fang Wu (2008). "Social networks that matter: Twitter under the microscope". In: *arXiv preprint arXiv:0812.1045*.

Hulovatyy, Yuriy and Tijana Milenković (2016). "SCOUT: simultaneous time segmentation and community detection in dynamic networks". In: *Scientific reports* 6. DOI: 10.1038/srep37557.

Hulsermann, Ralf, Andreas Betker, Monika Jager, Stefan Bodamer, Marc Barry, Jan Spath, Christoph Gauger, and Martin Kohn (2004). "A set of typical transport network scenarios for network modelling". In: *ITG Fachbericht* 182, pp. 65–72.

Jackson, Matthew O (2011). "An overview of social networks and economic applications". In: *Handbook of social economics*. Vol. 1. Elsevier, pp. 511–585.

Janson, Svante, Tomasz Luczak, and Andrzej Rucinski (2011). *Random graphs*. John Wiley & Sons.

Jin, Jiashun, Zheng Ke, and Shengming Luo (2018). "Network global testing by counting graphlets". In: *International Conference on Machine Learning*. PMLR, pp. 2333–2341.

Jin, Jiashun, Zheng Tracy Ke, and Shengming Luo (2021). "Optimal adaptivity of signed-polygon statistics for network testing". In: *The Annals of Statistics* 49.6, pp. 3408–3433. DOI: 10.1214/21-AOS2089.

Keane, Nathan, Connie Yee, and Liang Zhou (2015). "Using Topic Modeling and Similarity Thresholds to Detect Events". In: *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. ACL, pp. 34–42. DOI: 10.3115/v1/W15-0805.

Kendrick, Lucy, Katarzyna Musial, and Bogdan Gabrys (2018). "Change point detection in social networks: critical review with experiments". In: *Computer Science Review* 29, pp. 1–13. DOI: 10.1016/j.cosrev.2018.05.001.

Kim, Taehoon and Jaesik Choi (2015). "Reading documents for bayesian online change point detection". In: *Proceedings of the 2015 EMNLP-Conference*. ACL, pp. 1610–1619. DOI: 10.18653/v1/D15-1184.

Knittel, Johannes, Steffen Koch, and Thomas Ertl (2021). "PyramidTags: Context-, Time- and Word Order-Aware Tag Maps to Explore Large Document Collections". In: *IEEE Transactions on Visualization and Computer Graphics* 27.12, pp. 4455–4468. DOI: 10.1109/tvcg.2020.3010095.

Knoth, Sven (2017). "ARL numerics for MEWMA charts". In: *Journal of Quality Technology* 49.1, pp. 78–89. DOI: https://doi.org/10.1080/00224065.2017.11918186.

Knoth, Sven and Wolfgang Schmid (2004). "Control charts for time series: a review". In: *Frontiers in statistical quality control 7*. Springer, pp. 210–236.

Kolaczyk, Eric D and Gábor Csárdi (2014). *Statistical analysis of network data with R*. Vol. 65. Springer.

Koutra, Danai, Neil Shah, Joshua T Vogelstein, Brian Gallagher, and Christos Faloutsos (2016). "Deltacon: Principled massive-graph similarity function with attribution". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 10.3, pp. 1–43. DOI: https://doi.org/10.1145/2824443.

Krivitsky, Pavel N, Mark S Handcock, and Martina Morris (2011). "Adjusting for network size and composition effects in exponential-family random graph models". In: *Statistical methodology* 8.4, pp. 319–339. DOI: 10.1016/j.stamet.2011.01.005.

Krstajic, Milos, Enrico Bertini, and Daniel Keim (2011). "CloudLines: Compact Display of Event Episodes in Multiple Time-Series". In: *IEEE Transactions on Visualization and Computer Graphics* 17.12, pp. 2432–2439. DOI: 10.1109/tvcg.2011.179.

Lee, Alan J (2019). *U-statistics: Theory and Practice*. Routledge.

Lee, Der-Horng and Meng Dong (2009). "Dynamic network design for reverse logistics operations under uncertainty". In: *Transportation Research Part E: Logistics and Transportation Review* 45.1, pp. 61–71. DOI: 10.1016/j.tre.2008.08.002.

Lei, Jing (2016). "A goodness-of-fit test for stochastic block models". In: *The Annals of Statistics* 44.1, pp. 401–424. DOI: 10.1214/15-AOS1370.

Li, Yuemeng, Aidong Lu, Xintao Wu, and Shuhan Yuan (2019). "Dynamic anomaly detection using vector autoregressive model". In: *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part I 23*. Springer, pp. 600–611. DOI: 10.1007/978-3-030-16148-4_46.

Liang, Qiao and Kaibo Wang (2019). "Monitoring of user-generated reviews via a sequential reverse joint sentiment-topic model". In: *Quality and Reliability Engineering International* 35.4, pp. 1180–1199. DOI: 10.1002/qre.2452.

Lin, Chuan-hao, Linchuan Xu, and Kenji Yamanishi (2022). "Network Change Detection Based on Random Walk in Latent Space". In: *IEEE Transactions on Knowledge and Data Engineering*. DOI: 10.1109/TKDE.2022.3167062.

Liu, Shixia, Jialun Yin, Xiting Wang, Weiwei Cui, Kelei Cao, and Jian Pei (2016). "Online Visual Analytics of Text Streams". In: *IEEE Transactions on Visualization and Computer Graphics* 22.11, pp. 2451–2466. DOI: 10.1109/tvcg.2015.2509990.

Liu, Shixia, Michelle X. Zhou, Shimei Pan, Yangqiu Song, Weihong Qian, Weijia Cai, and Xiaoxiao Lian (2012). "TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis". In: *ACM Transactions on Intelligent Systems and Technology* 3.2, pp. 1–28. DOI: 10.1145/2089094.2089101.

Lospinoso, Josh and Tom AB Snijders (2019). "Goodness of fit for stochastic actor-oriented models". In: *Methodological Innovations* 12.3, p. 2059799119884282. DOI: 10.1177/2059799119884282.

Lovász, László (2012). *Large networks and graph limits*. Vol. 60. American Mathematical Soc.

Lowry, Cynthia A, William H Woodall, Charles W Champ, and Steven E Rigdon (1992). "A multivariate exponentially weighted moving average control chart". In: *Technometrics* 34.1, pp. 46–53. DOI: 10.1080/00401706.1992.10485232.

Lu, Yafeng, Hong Wang, Steven Landis, and Ross Maciejewski (2018). "A Visual Analytics Framework for Identifying Topic Drivers in Media Events". In: *IEEE Transactions on Visualization and Computer Graphics* 24.9, pp. 2501–2515. DOI: 10.1109/tvcg.2017.2752166.

Luo, Dongning, Jing Yang, Milos Krstajic, William Ribarsky, and Daniel Keim (2012). "EventRiver: Visually Exploring Text Collections with Temporal References". In: *IEEE Transactions on Visualization and Computer Graphics* 18, pp. 93–105. DOI: 10.1109/TVCG.2010.225.

Malinovskaya, Anna and Philipp Otto (2021). "Online network monitoring". In: *Statistical Methods & Applications* 30.5, pp. 1337–1364. DOI: 10.1007/s10260-021-00589-z.

Marcus, Adam, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller (2011). "TwitInfo: Aggregating and visualizing microblogs for event exploration". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 227–236. DOI: 10.1145/1978942.1978975.

Maugis, Pierre A, Sofia C Olhede, Carey E Priebe, and Patrick J Wolfe (2020). "Testing for equivalence of network distribution using subgraph counts". In: *Journal of Computational and Graphical Statistics* 29.3, pp. 455–465. DOI: 10.1080/10618600.2020.1736085.

McCulloh, Ian and Kathleen M Carley (2011). "Detecting change in longitudinal social networks". In: *Military Academy West Point NY Network Science Center (NSC)*.

Mei, Qiaozhu and ChengXiang Zhai (2005). "Discovering Evolutionary Theme Patterns from Text: An Exploration of Temporal Text Mining". In: *Proceedings of the 11th SIGKDD-Conference*. ACM, pp. 198–207. ISBN: 159593135X. DOI: 10.1145/1081870.1081895.

Montgomery, Douglas C (2007). *Introduction to statistical quality control*. John Wiley & Sons.

– (2012). *Statistical quality control*. Wiley Global Education.

Morales, Jose Andre, Areej Al-Bataineh, Shouhuai Xu, and Ravi Sandhu (2010). "Analyzing and exploiting network behaviors of malware". In: *International conference on security and privacy in communication systems*. Springer, pp. 20–34. DOI: 10.1007/978-3-642-16161-2_2.

Motalebi, Narges, Mohammad Saleh Owlia, Amirhossein Amiri, and Mohammad Saber Fallahnezhad (2021). "Monitoring social networks based on Zero-inflated Poisson regression model". In: *Communications in Statistics-Theory and Methods*, pp. 1–17. DOI: 10.1080/03610926.2021.1945103.

Neil, Joshua, Curtis Hash, Alexander Brugh, Mike Fisk, and Curtis B Storlie (2013). "Scan statistics for the online detection of locally anomalous subgraphs". In: *Technometrics* 55.4, pp. 403–414. DOI: 10.1080/00401706.2013.822830.

Newman, Mark (2018). *Networks*. Oxford university press.

Nguyen, Giang Hoang, John Boaz Lee, Ryan A Rossi, Nesreen K Ahmed, Eunyee Koh, and Sungchul Kim (2018). "Continuous-time dynamic network embeddings". In: *Companion proceedings of the the web conference 2018*, pp. 969–976. DOI: 10.1145/3184558.3191526.

Nowicki, Krzysztof (1989). "Asymptotic normality of graph statistics". In: *Journal of Statistical Planning and Inference* 21.2, pp. 209–222. DOI: 10.1016/0378-3758(89)90005-0.

Nowicki, Krzysztof and John C Wierman (1988). "Subgraph counts in random graphs using incomplete U-statistics methods". In: *Discrete Mathematics* 72.1-3, pp. 299–310. DOI: 10.1016/0012-365X(88)90220-8.

Oakland, Robert James and John S Oakland (2018). *Statistical process control*. Routledge.

Ofori-Boateng, Dorcas, Yulia R Gel, and Ivor Cribben (2021). "Nonparametric anomaly detection on time series of graphs". In: *Journal of Computational and Graphical Statistics* 30.3, pp. 756–767. DOI: 10.1080/10618600.2020.1844214.

Ospina-Forero, Luis, Charlotte M Deane, and Gesine Reinert (2019). "Assessment of model fit via network comparison methods based on subgraph counts". In: *Journal of Complex Networks* 7.2, pp. 226–253. DOI: 10.1093/comnet/cny017.

Ou, Mingdong, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu (2016). "Asymmetric transitivity preserving graph embedding". In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1105–1114. DOI: 10.1145/2939672.2939751.

Ouadah, Sarah, Stéphane Robin, and Pierre Latouche (2020). "Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases". In: *Scandinavian Journal of Statistics* 47.1, pp. 156–181. DOI: 10.1111/sjos.12410.

Park, Youngser, C Priebe, D Marchette, and Abdou Youssef (2009). "Anomaly detection using scan statistics on time series hypergraphs". In: *Link Analysis, Counterterrorism and Security (LACTS) Conference*. SIAM Pennsylvania, p. 9.

Pasquali, Arian, Ricardo Campos, Alexandre Ribeiro, Brenda Santana, Alípio Jorge, and Adam Jatowt (2021). "TLS-Covid19: A New Annotated Corpus for Timeline Summarization". In: *Advances in Information Retrieval, ECIR 2021*. Vol. 12656. LNCS, pp. 497–512. DOI: 10.1007/978-3-030-72113-8_33.

Paull, Sara H, Sejin Song, Katherine M McClure, Loren C Sackett, A Marm Kilpatrick, and Pieter TJ Johnson (2012). "From superspreaders to disease hotspots: linking transmission across hosts and space". In: *Frontiers in Ecology and the Environment* 10.2, pp. 75–82. DOI: 10.1890/110111.

Peel, Leto and Aaron Clauset (2015). "Detecting change points in the large-scale structure of evolving networks". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2914–2920. DOI: 10.1609/aaai.v29i1.9574.

Perozzi, Bryan, Rami Al-Rfou, and Steven Skiena (2014). "Deepwalk: Online learning of social representations". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710. DOI: https://doi.org/10.1145/2623330.2623732.

Phaladiganon, Poovich, Seoung Bum Kim, Victoria CP Chen, Jun-Geol Baek, and Sun-Kyoung Park (2011). "Bootstrap-based T2 multivariate control charts". In: *Communications in Statistics - Simulation and Computation* 40.5, pp. 645–662. DOI: 10.1080/03610918.2010.549989.

Pincombe, Brandon (2005). "Anomaly detection in time series of graphs using arma processes". In: *Asor Bulletin* 24.4, p. 2.

Prabhu, Sharad S and George C Runger (1997). "Designing a multivariate EWMA control chart". In: *Journal of Quality Technology* 29.1, pp. 8–15. DOI: 10.1080/00224065.1997.11979720.

Priebe, Carey E, John M Conroy, David J Marchette, and Youngser Park (2005). "Scan statistics on enron graphs". In: *Computational & Mathematical Organization Theory* 11.3, pp. 229–247. DOI: 10.1007/s10588-005-5378-z.

Prill, Robert J, Pablo A Iglesias, and Andre Levchenko (2005). "Dynamic properties of network motifs contribute to biological network organization". In: *PLoS biology* 3.11, e343. DOI: 10.1371/journal.pbio.0030343.

Ranshous, Stephen, Shitian Shen, Danai Koutra, Steve Harenberg, Christos Faloutsos, and Nagiza F Samatova (2015). "Anomaly detection in dynamic networks: a survey". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 7.3, pp. 223–247. DOI: 10.1002/wics.1347.

Rieger, Jonas, Carsten Jentsch, and Jörg Rahnenführer (2021). "RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data". In: *Findings Proceedings of the 2021 EMNLP-Conference*. ACL, pp. 2337–2347. DOI: 10.18653/v1/2021.findings-emnlp.201.

Rieger, Jonas, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch (2022). "Dynamic change detection in topics based on rolling LDAs". In: *Proceedings of the Text2Story'22 Workshop*. Vol. 3117. CEUR-WS, pp. 5–13. URL: http://ceur-ws.org/Vol-3117/.

Rissanen, Jorma (1989). *Stochastic complexity in statistical inquiry*. Vol. 15. World Scientific.

Roberts, Stuart W (1959). "Control Chart Tests Based on Geometric Moving Averages". In: *Technometrics* 1.3, pp. 239–250. DOI: 10.1080/00401706.2000.10485986.

Robins, Garry, Pip Pattison, Yuval Kalish, and Dean Lusher (2007). "An introduction to exponential random graph (p*) models for social networks". In: *Social Networks* 29.2, pp. 173–191. DOI: 10.1016/j.socnet.2006.08.002.

Rodriguez-Nunez, Eduardo and Juan Carlos Garcia-Palomares (2014). "Measuring the vulnerability of public transport networks". In: *Journal of transport geography* 35, pp. 50–63. DOI: 10.1016/j.jtrangeo.2014.01.008.

Salmasnia, Ali, Mohammadreza Mohabbati, and Mohammadreza Namdar (2020). "Change point detection in social networks using a multivariate exponentially weighted moving average chart". In: *Journal of Information Science* 46.6, pp. 790–809. DOI: https://doi.org/10.1177/0165551519863351.

Sarkar, Purnamrita and Andrew W Moore (2005). "Dynamic social network analysis using latent space models". In: *Acm Sigkdd Explorations Newsletter* 7.2, pp. 31–40. DOI: 10.1145/1117454.111745.

Snijders, Tom AB (1981). "The degree variance: an index of graph heterogeneity". In: *Social Networks* 3.3, pp. 163–174. DOI: 10.1016/0378-8733(81)90014-9.

– (1996). "Stochastic actor-oriented models for network change". In: *Journal of Mathematical Sociology* 21.1-2, pp. 149–172. DOI: 10.1080/0022250X.1996.9990178.

Song, Xiaodan, Ching-Yung Lin, Belle L. Tseng, and Ming-Ting Sun (2005). "Modeling and Predicting Personal Information Dissemination Behavior". In: *Proceedings of the 11th SIGKDD-Conference*. ACM, pp. 479–488. DOI: 10.1145/1081870.1081925.

Stoumbos, Zachary G and Joe H Sullivan (2002). "Robustness to non-normality of the multivariate EWMA control chart". In: *Journal of Quality Technology* 34.3, pp. 260–276. DOI: 10.1080/00224065.2002.11980157.

Sun, Guodao, Yingcai Wu, Shixia Liu, Tai-Quan Peng, Jonathan J. H. Zhu, and Ronghua Liang (2014). "EvoRiver: Visual Analysis of Topic Coopetition on Social Media". In: *IEEE Transactions on Visualization and Computer Graphics* 20.12, pp. 1753–1762. DOI: 10.1109/tvcg.2014.2346919.

Sun, Jimeng, Dacheng Tao, and Christos Faloutsos (2006). "Beyond streams and graphs: dynamic tensor analysis". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 374–383. DOI: 10.1145/1150402.1150445.

Sun, Tong and Yan Liu (2018). "A dynamic network change detection method using network embedding". In: *Cloud Computing and Security: 4th International Conference, ICCCS 2018, Haikou, China, June 8-10, 2018, Revised Selected Papers, Part I 4*. Springer, pp. 63–74. DOI: 10.1007/978-3-030-00006-6_6.

Vanhems, Philippe, Alain Barrat, Ciro Cattuto, Jean-François Pinton, Nagham Khanafer, Corinne Régis, Byeul-a Kim, Brigitte Comte, and Nicolas Voirin (2013). "Estimating potential infection transmission routes in hospital wards using wearable proximity sensors". In: *PloS one* 8.9. DOI: 10.1371/journal.pone.0073970.

Wang, Chong, David M. Blei, and David Heckerman (2008). "Continuous Time Dynamic Topic Models". In: *Proceedings of the 24th UAI-Conference*. AUAI Press, pp. 579–586. URL: https://dl.acm.org/doi/10.5555/3023476.3023545.

Wang, Jiaoe, Huihui Mo, Fahui Wang, and Fengjun Jin (2011). "Exploring the network structure and nodal centrality of China's air transport network: A complex network approach". In: *Journal of Transport Geography* 19.4, pp. 712–721. DOI: 10.1016/j.jtrangeo.2010.08.012.

Wang, Xuerui and Andrew McCallum (2006). "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends". In: *Proceedings of the 12th SIGKDD-Conference*. ACM, pp. 424–433. DOI: 10.1145/1150402.1150450.

Wang, Yunli and Cyril Goutte (2018). "Real-time Change Point Detection using On-line Topic Models". In: *Proceedings of the 27th ACL-Conference*. ACL, pp. 2505–2515. URL: https://www.aclweb.org/anthology/C18-1212.

Watts, Duncan J and Steven H Strogatz (1998). "Collective dynamics of small-world networks". In: *nature* 393.6684, pp. 440–442.

Wen, Lei, Yachao Shi, and Lijun Wang (2013). "The modeling and simulation of supply chain based on directed complex network". In: *Journal of Information & Computational Science* 10.18, pp. 6085–6092.

Wilson, James D, Nathaniel T Stevens, and William H Woodall (2019). "Modeling and detecting change in temporal networks via the degree corrected stochastic block model". In: *Quality and Reliability Engineering International* 35.5, pp. 1363–1378. DOI: 10.1002/qre.2520.

Wu, Xinyu, Zhou Huang, Xia Peng, Yiran Chen, and Yu Liu (2018). "Building a spatially-embedded network of tourism hotspots from geotagged social media data". In: *IEEE Access* 6, pp. 21945–21955. DOI: 10.1109/ACCESS.2018.2828032.

Xie, Yingjie, Wenjun Wang, Minglai Shao, Tianpeng Li, and Yandong Yu (2023). "Multi-view change point detection in dynamic networks". In: *Information Sciences* 629, pp. 344–357. DOI: 10.1016/j.ins.2023.01.118.

Xu, Panpan, Yingcai Wu, Enxun Wei, Tai Quan Peng, Shixia Liu, Jonathan J.H. Zhu, and Huamin Qu (2013). "Visual analysis of topic competition on social media". In: *IEEE Transactions on Visualization and Computer Graphics* 19, pp. 2012–2021. DOI: 10.1109/TVCG.2013.221.

Xu, Wenkai and Gesine Reinert (2021). "A Stein goodness-of-test for exponential random graph models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 415–423.

Yu, Lisha, Inez Maria Zwetsloot, Nathaniel Tyler Stevens, James David Wilson, and Kwok Leung Tsui (2022). "Monitoring dynamic networks: A simulation-based strategy for comparing monitoring methods and a comparative study". In: *Quality and Reliability Engineering International* 38.3, pp. 1226–1250. DOI: 10.1002/qre.2944.

Yuan, Mingao, Ruiqi Liu, Yang Feng, and Zuofeng Shang (2022). "Testing community structure for hypergraphs". In: *The Annals of Statistics* 50.1, pp. 147–169. DOI: 10.1214/21-AOS2099.

Zhang, Changhong, Zeyu Li, and Jiawan Zhang (2017). "A Survey on Visualization for Scientific Literature Topics". In: *Journal of Visualization* 21.2, pp. 321–335. DOI: 10.1007/s12650-017-0462-2.

Zhou, Lekui, Yang Yang, Xiang Ren, Fei Wu, and Yueting Zhuang (2018). "Dynamic network embedding by modeling triadic closure process". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. DOI: 10.1609/aaai.v32i1.11257.

Zou, Na and Jing Li (2017). "Modeling and change detection of dynamic network data by a network state space model". In: *IISE Transactions* 49.1, pp. 45–57. DOI: 10.1080/0740817X.2016.1198065.