technische universität
dortmund

Clustering Malware for Generating Behavioral Signatures

Martin Apel, Michael Meier
Email: {martin.apel,michael.meier}@udo.edu

Technische Universität Dortmund

7. Juli 2010

Überblick

1. Clusterverfahren
2. Clustern zur Signaturerstellung
3. Kriterien und Experimente
4. Ergebnisse
5. Auswertung und Ausblick

Gruppierung

Ansätze bzw. Techniken zur Gruppierung sind:

- Verwendung von verhaltensbasierter Analyse
- Verwendung von Cluster Algorithmen, um ähnliches Verhalten zu gruppieren
- Verwandte Ansätze:
  - Automated classification and analysis of internet malware, Bailey et al.
  - Scalable, Behaviour based malware clustering, Bayer et al.
  - Automatic Analysis of Malware Behaviour using Machine Learning, Rieck et al.
- Fokus unserer Arbeit: Signaturgenerierung

Cluster Algorithmen

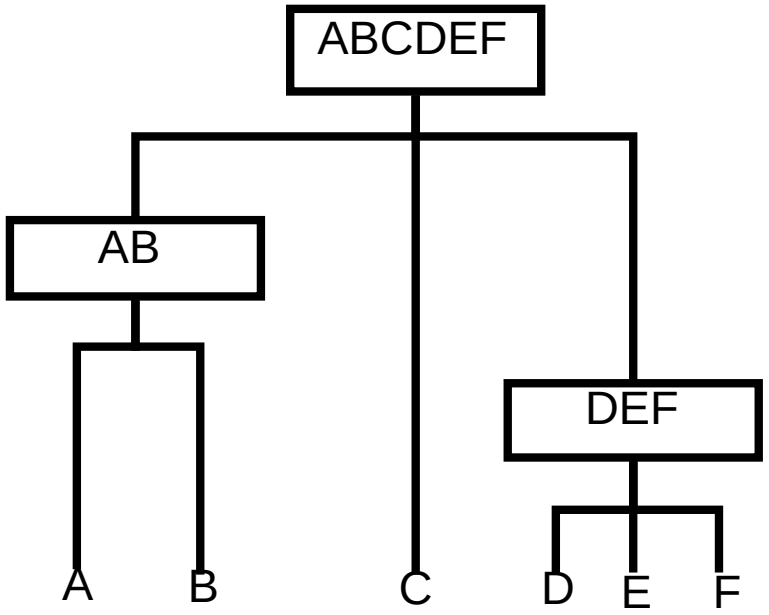
- Eingabe: Menge von Objekten M
- Ausgabe: Gruppierung der Eingabeobjekte M
- Gruppierung geschieht an Hand von Eigenschaften der Objekte (Abstandsfunktion)
- Varianten von Clusterverfahren:
  - inkrementell
  - überlappend
  - hierarchisch

Hierarchische Cluster Algorithmen

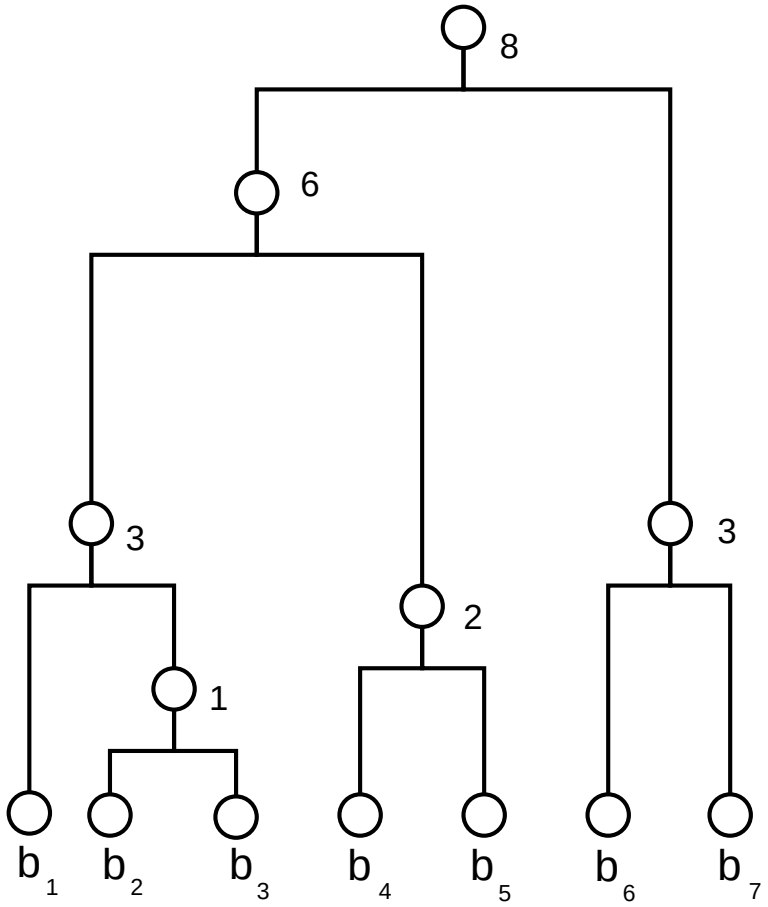
Algorithmus:

1. Für jedes zu clusternde Element wird ein Cluster mit diesem Element erzeugt.
2. So lang die jeweils ähnlichsten Cluster zu neuen Clustern zusammenfassen, bis nur noch ein Cluster vorhanden ist.

Ausgabe: Hierarchie von Clustern



Hierarchische Cluster Algorithmen und Signaturgenerierung



- Signaturkandidaten für jeden Cluster mit mindestens zwei Elementen ermitteln
- Qualitätskontrolle der Signatur-kandidaten (Goodpoolcheck)
- Auswahl einer überdeckenden Menge von Signaturen

Varianten hierarchischer Cluster Algorithmen

Single-Link

- maximiert den Abstand zwischen den Clustern (Inter-Cluster Abstand)
- neigt zur Kettenbildung

Complete-Link

- minimiert Abstand innerhalb der Cluster (Diversität)
- die entstehenden Cluster sind häufig relativ klein

WPGMA und UPGMA

- finden durch Mittlung einen Kompromiss zwischen dem Maximieren des Inter-Cluster Abstands und der Minimierung der Diversität
- WPGMA: Später einbezogene Elemente werden stärker gewichtet
- UPGMA: Alle Elemente werden gleich gewichtet

Welches Clusterverfahren verwenden?

- Vermutung: Geringe Diversität ist wichtig für die Signaturgenerierung, da hierbei Gemeinsamkeiten fokussiert werden.
  $\Rightarrow$ Complete-Link
- Experimentelle Überprüfung an synthetischen und realen Daten

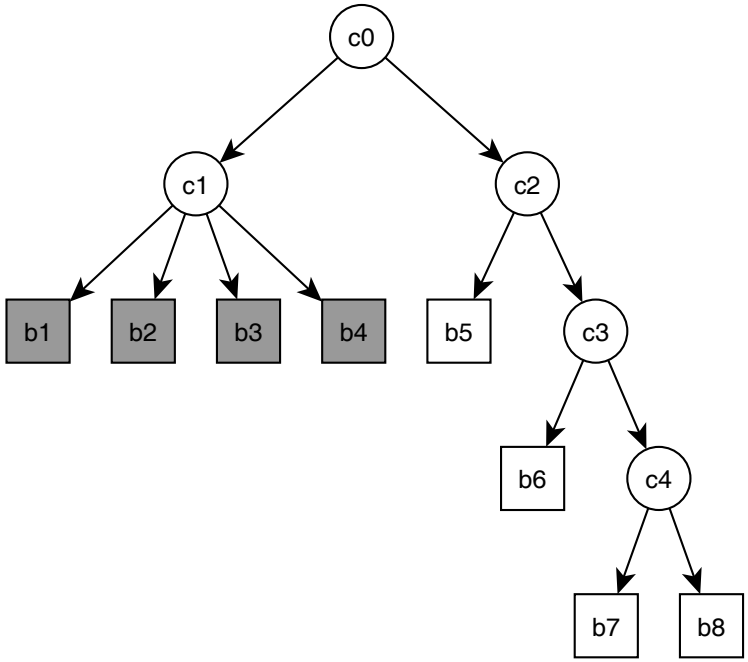
Kriterien - Synthetisches Datenset

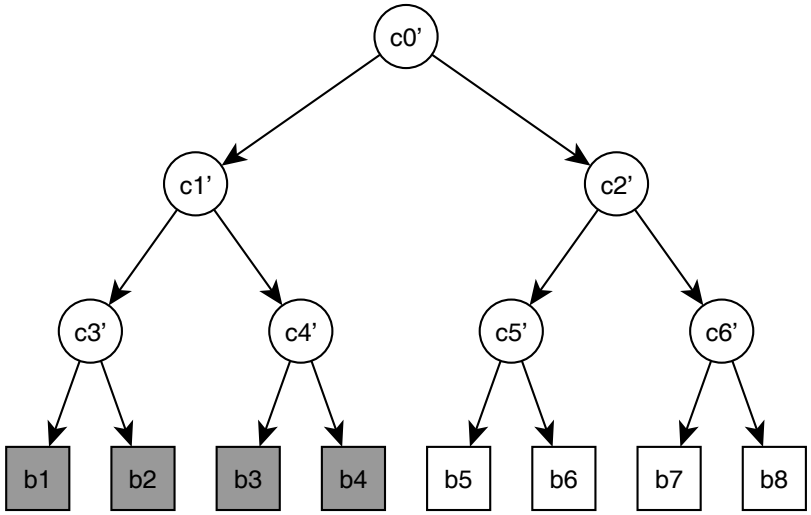
- Werden die folgenden Ähnlichkeitsrelationen korrekt priorisiert:

Relation	Beispiel für ähnliche Objekte
$=_0$	$A =_0 A$
$=_1$	$xA\hat{x} = yA\hat{y}$
$=_2$	$xA_1...\hat{x}A_n =_2 yA_1...\hat{y}A_n$
$=_3$	$xA_1...\hat{x}xA_n =_3 yA_n...\hat{y}A_1$

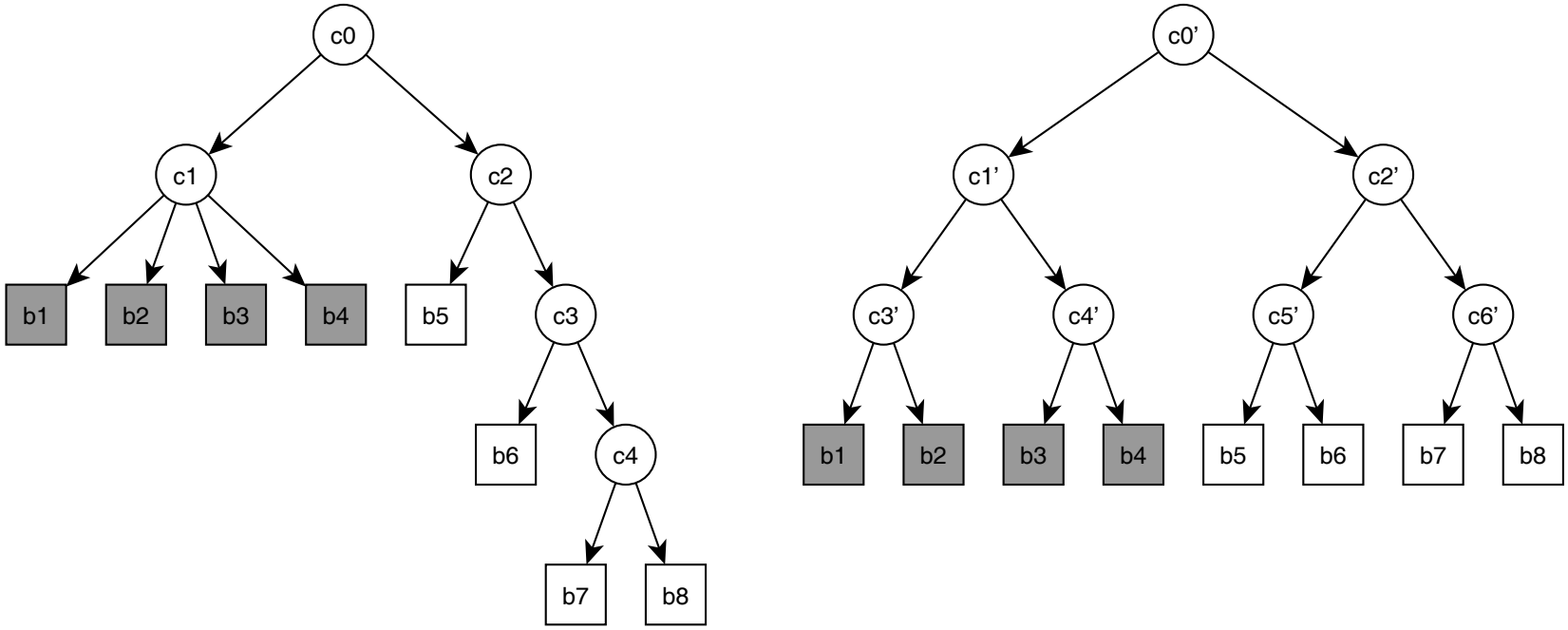
- Verhalten bei verrauschten Eingabedaten. Sind die entstehenden Clusterhierarchien
  - korrekt im Sinne der Priorisierung der Ähnlichkeitsrelationen?
  - balanciert?

Warum Balance?





Warum Balance?



- Verwendetes Maß für Balance: Standard Abweichung der Blatttiefen

Kriterien - Realdatenset

- Kein zuverlässiger Ground-Truth vorhanden, daher nur *Vergleich* anhand der Ergebnisse (Signaturen) möglich
- Unser Fokus ist Signaturgenerierung
  $\Rightarrow$ Qualität mit Hilfe von Signaturen testen
- Signaturerstellung mit stark vereinfachter Qualitätsbedingung: Signaturkandidat ist gut genug, wenn er mindestens 28 Systemrufe enthält
- Vergleich zweier Mengen von Signaturen A und B:
  - Gilt $|V(A)| > |V(B)|$ so ist A besser als B
  - Gilt $|V(A)| = |V(B)|$ und $|A| < |B|$ so ist A besser als B

Aufbau der Datensets

Synthetisches Datenset

- Für Priorisierungstest: Je ein Datenset für jede zu testende Relation.
- Für Tests mit verrauschten Eingabedaten:
  Je 50 Sets für jede zu testende Menge an Rauschen.

Realdatenset

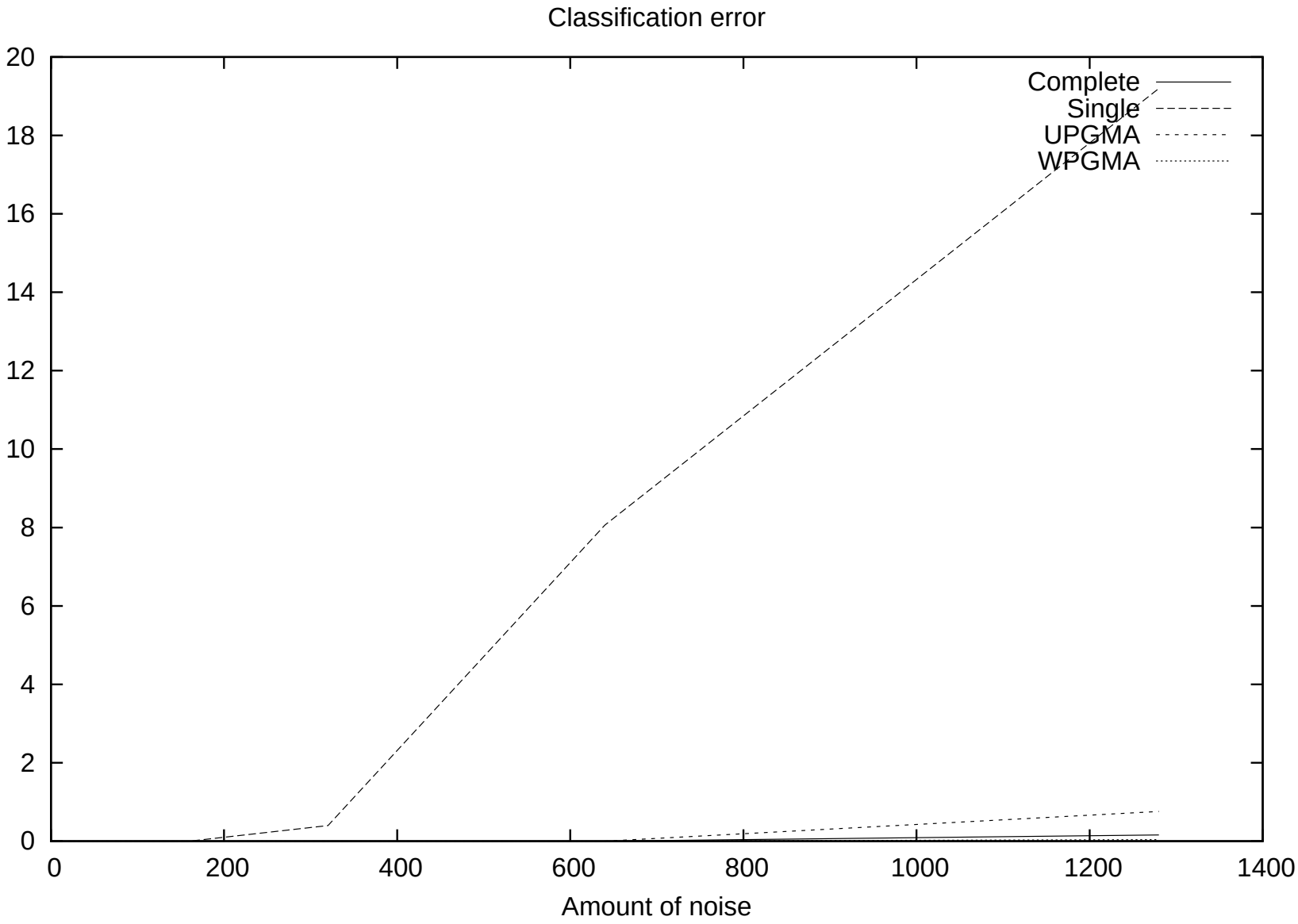
- Berichte von 1623 Samples
- Jedes Sample hat mindestens einen Partner mit dem es sich Verhalten teilt (min. 28 Api-Calls)

Ergebnis - Priorisierung der Relationen

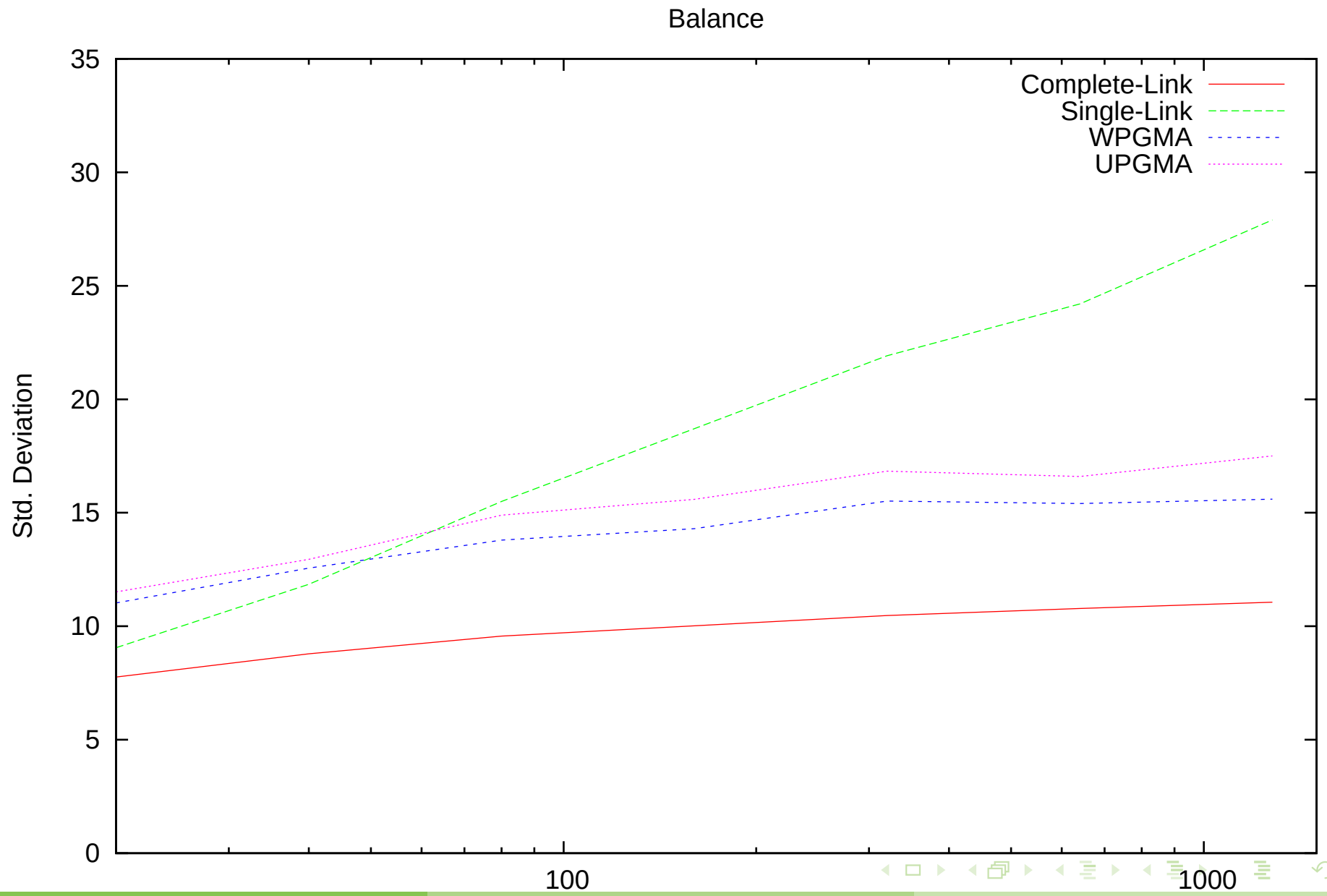
Priorisierung der Relationen ist bei allen Algorithmen gleich:

$$=_0 \text{ vor } =_1 \text{ vor } =_2 \text{ wie } =_3$$

Ergebnis - Klassifikationsfehler bei verrauschten Daten



Ergebnis - Balance der Hierarchie bei verrauschten Daten



Ergebnis - Abdeckung des Realdatensets

Algorithmus	Abdeckung	Clusteranzahl	Std-Dev. der Blatttiefen
Single-Link	95.17%	66	32.7
Complete-Link	99.21%	51	5.6
WPGMA	99%	56	8.4
UPGMA	98.42%	58	15.1

- Complete-Link liefert die beste Abdeckung und benötigt hierfür am wenigsten Cluster.

Zusammenfassung

Algorithmus	Priori-sierung	Klassifikations-fehler	Balance	Abdeckung
Single-Link	+	−	−	95.17%
Complete-Link	+	+	++	99.21%
WPGMA	+	+	+	99%
UPGMA	+	+	+	98.42%

Stand im Amselprojekt

Zwei Clusterverfahren werden verwendet:

- Statisches Clustering (Complete-Link)
- Inkrementelles und überlappendes Clustering

Ausblick

- Verwendete Abstandsfunktion bzw. Abstandsmatrix ist problematisch
- Ähnlichkeit zum gelernten Konzept (Signatur) wichtiger als Abstand zum Cluster
  $\Rightarrow$ Conceptual Clustering

Vielen Dank für die Aufmerksamkeit

Fragen?

Inkrementelles Clustering im Amsel Projekt