# Linkbasierte Strukturierung von Web-Seiten

## Konzepte und Methoden

Dr. Michael Brinkmeier

Institut für Theoretische Informatik
Technische Universität Ilmenau

28. September 2006

# Welche Informationen sind im WWW enthalten?

## Binsenweisheiten

- Das WWW ist eine der größten und wichtigsten Informationsquellen.
- Das WWW entwickelt sich unkontrolliert und damit sind die Informationen unstrukturiert.

## Kernfrage

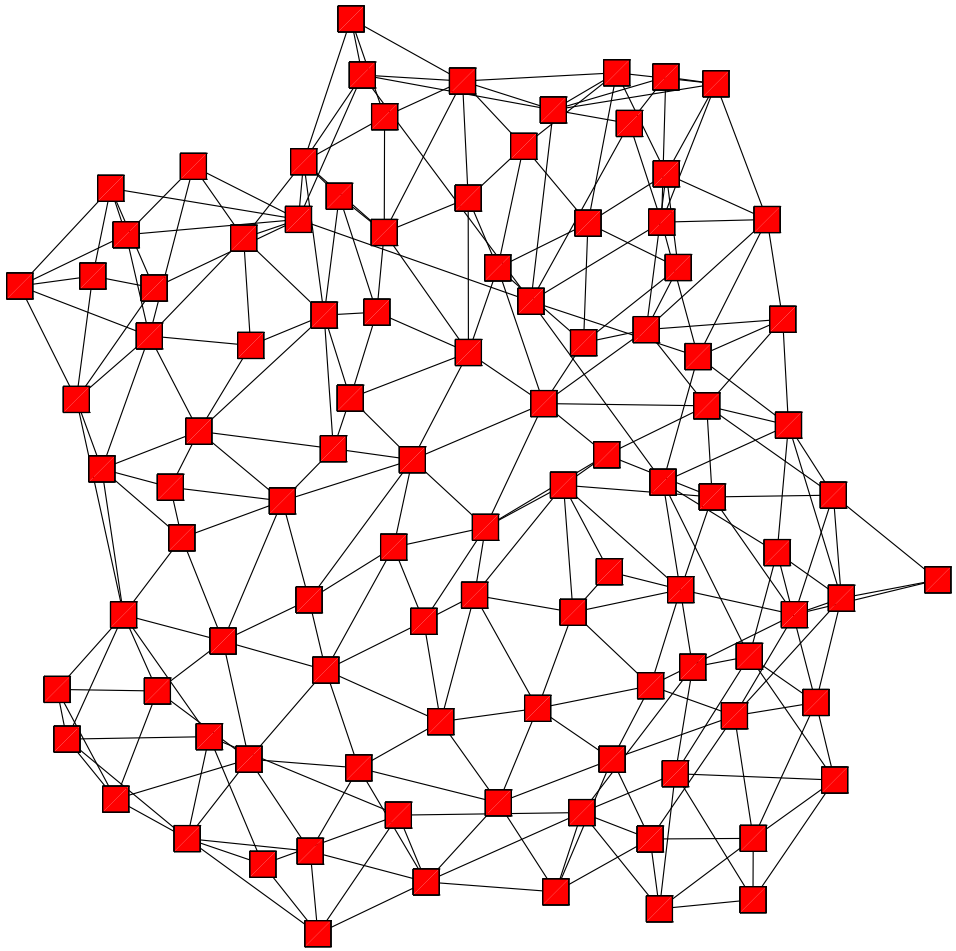
Welche Informationen sind im WWW enthalten und wie können sie extrahiert/handhabbar gemacht werden?

# Welche Informationen sind im WWW enthalten?

- **Inhalte (textuell und multimedial)**
  - „klassischer“ Ansatz $\Rightarrow$ Texte
  - Textbasierte Suchmaschine $\hat{=}$ Reverse Word Index,
- In Zukunft: **Semantische Informationen (Semantic Web)**
- **Links**
  - Bisher im Wesentlichen für das Ranking benutzt (z.B. PageRank, HITS)
  - Links werden von den Autoren **bewußt** gesetzt.
  - Ein Link stellt eine Beziehung zwischen zwei Seiten her
    - ★ inhaltlicher Zusammenhang
    - ★ soziale Beziehung (Freunde, Bekannte, Kollegen ...)
    - ★ wirtschaftliche Beziehung (Partnerfirmen, Kunden, Lieferanten ...)
    - ★ ...

# Mathematische Formalisierung

**Netzwerk/Graph** aus **Knoten** (Dokumente) und **Kanten** (Links)



# Kohäsive Gruppen

## Grundidee

Finde Gruppen von Knoten, die in irgendeiner Form dichter miteinander verbunden sind.

**Kohäsive Gruppen** oder **Communities**

- Der Begriff kohäsive (Unter-)Gruppe stammt aus der Soziometrie, genauer aus der Analyse sozialer Netzwerke
  - Knoten = Akteure / Individuen
  - Kanten = direkte Beziehungen
- Eine Gruppe von Individuen heißt kohäsiv, wenn die Beziehungen zwischen ihnen im Vergleich zum gesamten Netzwerk in irgendeinem Sinn relativ stark, direkt, intensiv, häufig oder positiv sind.

# Kohäsive Gruppen

- Verschiedenste (mathematische) Definitionen, jede mit spezifischen Vor- und Nachteilen

  (z.B. Cliquen, Kerne, LS-Mengen, $\lambda$-Mengen, bipartite Kerne, FLG-Communities, ...)

- Problem 1: (Effiziente) Berechnung dieser Gruppen
  - Algorithmen
  - verteilte Berechnung
  - Datenstrukturen
- Problem 2: sehr große Netzwerke
  - Speicherplatzintensiv
  - Entschärfung durch verteilte Berechnung möglich?

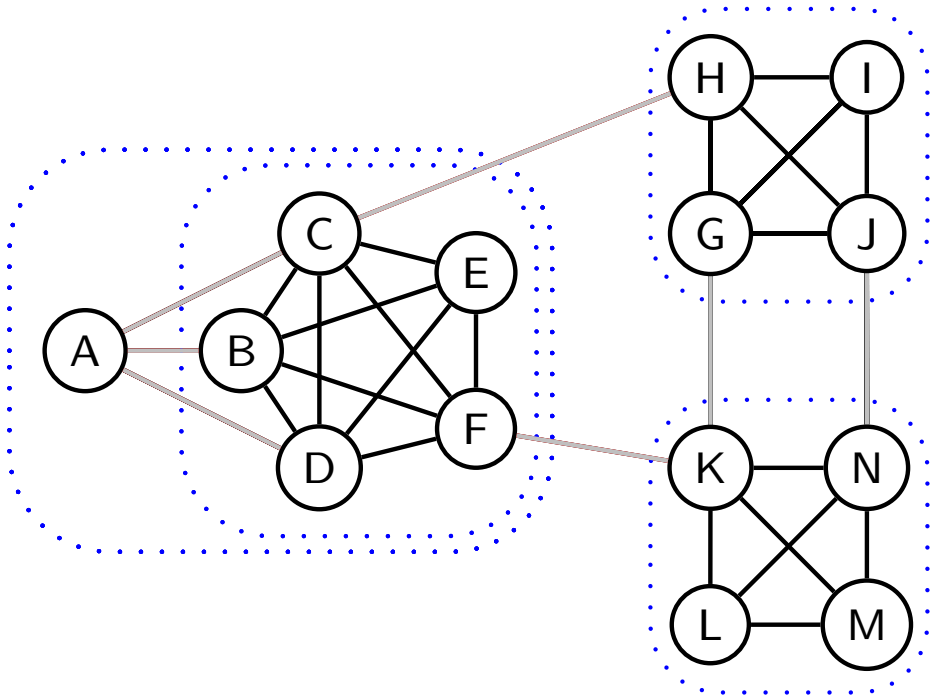
# Nutzung von kohäsiven Gruppen im WWW

- Nicht nur reine Auflistung der gefundenen Seiten nach Relevanz
- Gruppierung der Ergebnisse nach Communities
- Zusätzliche Anzeige von besonders relevanten Seiten, die sich in den gleichen Communities befinden (enthalten nicht unbedingt genau den gesuchten Begriff)
- Durchsuchen des WWW durch „Zusammenfalten“ von uninteressanten Gruppen oder Untergruppen

# Der von uns verwendete Begriff der Kohäsion

## Kohäsion = Kantenzusammenhang

Kohäsion = minimale Anzahl von Kanten, die entfernt werden müssen, damit die Gruppe in (mindestens) zwei Teile zerfällt.

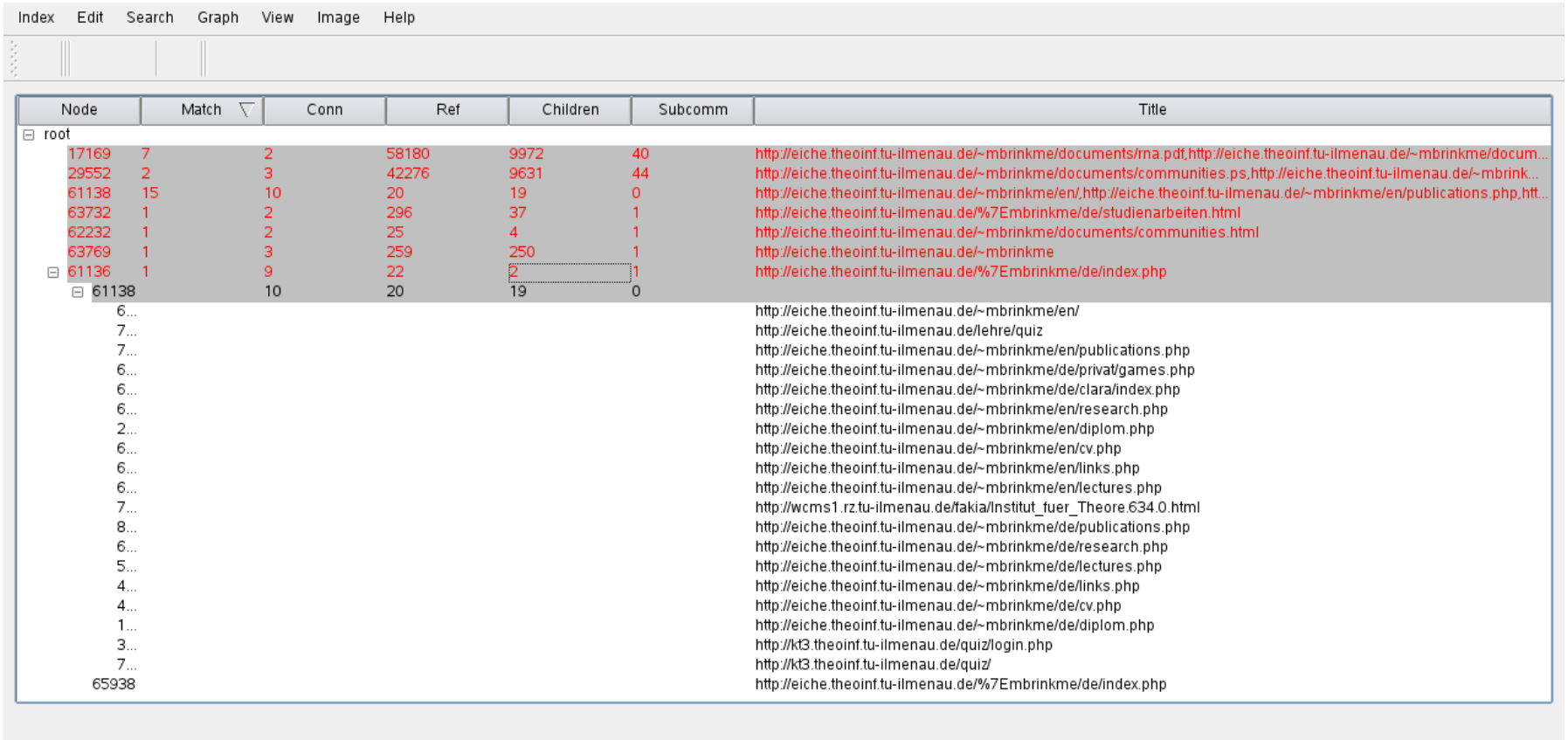



Diese Art der Kohäsion führt zu einer Hierarchie von Communities.

# Die Web-Domain tu-ilmenau.de

72.907 Knoten, 434.625 Kanten, 0:37:28 Stunden

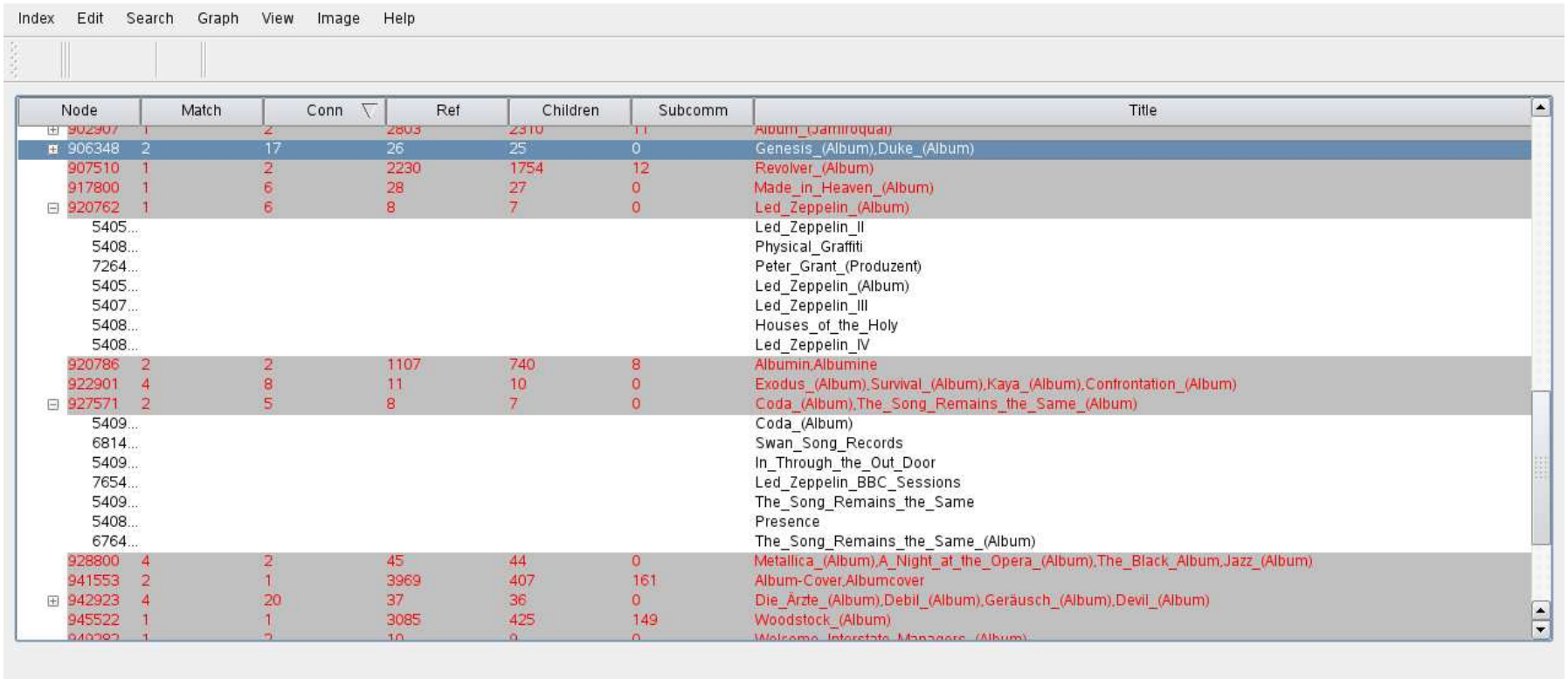
Beispiel: Suche nach mbrinkme in der URL



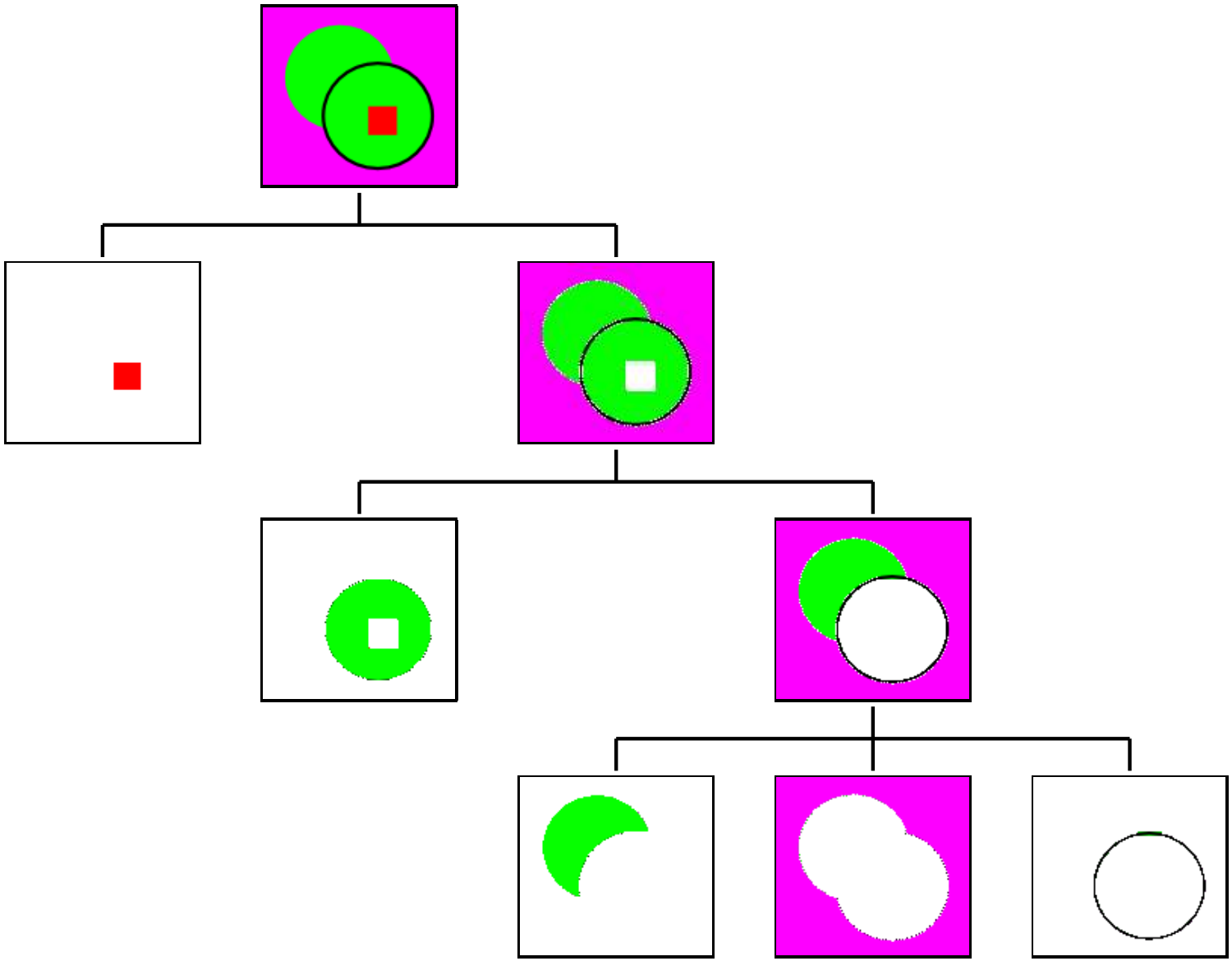
# Die deutsche Wikipedia

978.888 Knoten, 12.850.009 Kanten, ca. 4 Stunden

Beispiel: Suche nach `Album` im Titel des Eintrags



# Ein Bild sagt mehr als tausend Worte



# Aktuelle und zukünftige Arbeiten

- Verbesserung der Algorithmen/Implementierung
- Verbesserung der Vor- und Nachverarbeitung
- Abfrageschnittstelle
- Visualisierung
  - Layout der Netzwerke auf der Basis der Communities
- Erweiterung des Konzeptes auf sog. Hypergraphen
  - Wort-Dokument-Beziehungen
- Experimente mit verschiedenen Netzwerken

# Danke für Ihre Aufmerksamkeit!

Dr. Michael Brinkmeier

Institut für Theoretische Informatik
Technische Universität Ilmenau

mbrinkme@tu-ilmenau.de