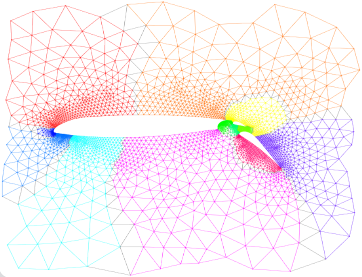
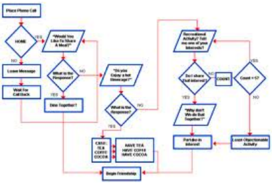
Karlsruher Institut für Technologie

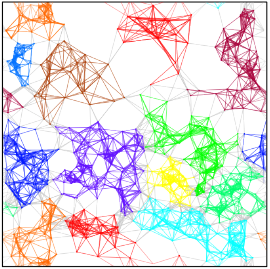
Algorithmische Methoden zur Netzwerkanalyse

Prof. Dr. Henning Meyerhenke

Institut für Theoretische Informatik







Vorlesung 2, Wiederholung und heutiges Thema

- Knotengradverteilung:
  “Wie viele Leute kenne ich?”
- *k*-Kerne:
  “Wie gut bin ich vernetzt?”

Vorlesung 2, Wiederholung und heutiges Thema

- Knotengradverteilung:
  “Wie viele Leute kenne ich?”
- *k*-Kerne:
  “Wie gut bin ich vernetzt?”
- **Heute auch:** Clusterkoeffizienten:
  “Der Freund meines Freundes ist mein Freund”
- NetworKit-Grundlagen

Inhalt

Struktur der Kernzerlegung

Obere und untere Schranken

Lemma

- *Sei* $G = (V, E)$ *ein Graph,*
- $(V_0, \ldots, V_k)$ *seine Kern-Zerlegung und*
- $G_i := (V_i, E_i) := G[V_i]$ *sein* i*-Kern.*

- *Sei* $n_i := |V_i \setminus V_{i+1}|$ *die Anzahl der Knoten in der* i*-Schale.*
- *Sei* $m_i := |E_i \setminus E_{i+1}|$ *die Anzahl der Kanten, die zu einem Knoten mit Kernzahl* i *und zu einem Knoten mit Kernzahl* $j \geq i$ *inzident sind.*
- *Per Definition:* $V_{k+1} := \emptyset$, $E_{k+1} := \emptyset$.

Struktur der Kernzerlegung

Obere und untere Schranken

Lemma

- *Sei* $n_i := |V_i \setminus V_{i+1}|$ *die Anzahl der Knoten in der* i*-Schale.*
- *Sei* $m_i := |E_i \setminus E_{i+1}|$ *die Anzahl der Kanten, die zu einem Knoten mit Kernzahl* i *und zu einem Knoten mit Kernzahl* $j \geq i$ *inzident sind.*
- *Per Definition:* $V_{k+1} := \emptyset$, $E_{k+1} := \emptyset$.

Dann lässt sich die Größe der i*-Schale beschränken durch:*

$$0 \leq n_i \leq |V| \tag{1}$$

$$\left.\begin{array}{ll} \lceil \frac{i \cdot n_i}{2} \rceil & \text{, if } n_i > i \\ \binom{n_i}{2} + n_i(i - n_i + 1) & \text{, if } n_i \leq i \end{array}\right\} \leq m_i \leq \begin{cases} i \cdot n_i & \text{, if } i < k \\ i \cdot n_i - \frac{i(i+1)}{2} & \text{, if } i = k \end{cases} \tag{2}$$

Beweis: Siehe Tafel!

Algorithmus zur Berechnung der k-Kern-Zerlegung

Algorithm 1 Berechnung der Kernzerlegung eines Graphen

function CoreDecomposition($G = (V, E)$)
  Output: Kernzahl k des Kerns von G und Feld *core* mit Kernzahl jedes Knotens
  Speichere die Knotengrade für alle Knoten in *resdegree*
  $i \leftarrow 0$
  while $V \neq \emptyset$ **do**
    $i \leftarrow i + 1$
    for each $v \in V$ mit $resdegree[v] < i$ **do**
      $core[v] \leftarrow i - 1$
      for each $u \in N(v)$ **do**
        $resdegree[u] \leftarrow resdegree[u] - 1$
      end for
      Entferne v aus G
    end for
  end while
  return $(i - 1, core)$
end function

Beispiel
[Baur et al., 2008]

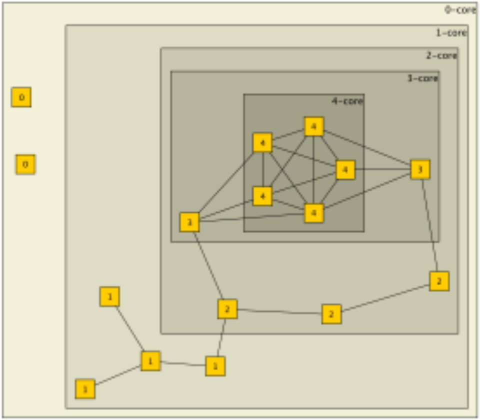


Fig. 1. A *k*-core decomposition with 5 core shells.

Laufzeit

Theorem

Sei $G = (V, E)$ mit $|E| = m$ zusammenhängend und ungewichtet.

Der Algorithmus CoreDecomposition *kann in*

Laufzeit $\mathcal{O}(m)$

implementiert werden.

Beweis: Siehe Tafel und Skript!

Parallele Berechung der Kernzerlegung

- Bräuchten eine parallele PQ
- Wird durch parallele Suche nach Knoten mit passendem Residualgraph ersetzt
- Diese Knoten und deren Nachbarschaft werden parallel abgearbeitet
- Siehe Implementierung in NetworKit

Literaturhinweise

N.S. Dasari, D. Ranjan, and M. Zubair: Park: An efficient algorithm for k-core decomposition on multicore processors. In Proc. Intl. Conference on Big Data 2014, pages 9–16.

Fazit *k*-Kern-Zerlegung

- Die *k*-Kernstruktur gibt ein Maß an, das robuster ist als der Knotengrad zur Bestimmung der Verbindung eines Knotens zum Rest des Graphen
- Lemmas beleuchten Größe der *i*-Schalen
- Algorithmische Definition führt zu Algorithmus, der die Struktur berechnet

Inhalt

Cluster-Koeffizient
Einführung

Fragestellung

- Ihnen gibt jemand einen Graphen und behauptet, es sei ein soziales Netzwerk.
- Sie möchten die Behauptung überprüfen.
- Welche Analyse-Möglichkeiten fallen Ihnen dazu bereits ein?

Cluster-Koeffizient

Einführung

Fragestellung

- Ihnen gibt jemand einen Graphen und behauptet, es sei ein soziales Netzwerk.
- Sie möchten die Behauptung überprüfen.
- Welche Analyse-Möglichkeiten fallen Ihnen dazu bereits ein?

Hinweis: Erwarten Sie relativ viele oder relativ wenige Dreiecke im Netzwerk?

Cluster-Koeffizient

Einführung

Fragestellung

- Ihnen gibt jemand einen Graphen und behauptet, es sei ein soziales Netzwerk.
- Sie möchten die Behauptung überprüfen.
- Welche Analyse-Möglichkeiten fallen Ihnen dazu bereits ein?

Hinweis: Erwarten Sie relativ viele oder relativ wenige Dreiecke im Netzwerk?

Aufgabe: Formalisierung dieses Konzeptes

Cluster-Koeffizient

Beschreibung

- Der Cluster-Koeffizient misst die durchschnittliche Wahrscheinlichkeit, dass zwei Nachbarn desselben Knotens zueinander benachbart sind.

$\Rightarrow$ Maß für die Dichte von Dreiecken im Netzwerk

Cluster-Koeffizient

Beschreibung

- Der Cluster-Koeffizient misst die durchschnittliche Wahrscheinlichkeit, dass zwei Nachbarn desselben Knotens zueinander benachbart sind.

$\Rightarrow$ Maß für die Dichte von Dreiecken im Netzwerk

- Gleiche Netzwerktypen haben häufig ähnliche Koeffizienten
- Unterschiedliche Netzwerktypen haben häufig unterschiedliche Koeffizienten
- Aber: Ausnahmen nicht so selten

Cluster-Koeffizient

Beschreibung

- Der Cluster-Koeffizient misst die durchschnittliche Wahrscheinlichkeit, dass zwei Nachbarn desselben Knotens zueinander benachbart sind.

$\Rightarrow$ Maß für die Dichte von Dreiecken im Netzwerk

- Gleiche Netzwerktypen haben häufig ähnliche Koeffizienten
- Unterschiedliche Netzwerktypen haben häufig unterschiedliche Koeffizienten
- Aber: Ausnahmen nicht so selten

- Beispiel in [Newman, S. 237]: Werte oft zwischen 0.1 und 0.6

$\Rightarrow$ Typische Wkt. dafür, dass gemeinsame Nachbarn zueinander benachbart sind, liegt zwischen 10% und 60%.

Cluster-Koeffizient

Definition

Definition

Der globale Cluster-Koeffizient C ist definiert als

$$C := \frac{3 \cdot \text{Zahl der (geschlossenen) Dreiecke}}{\text{Zahl der zusammenhängenden Dreiergruppen}}$$

Cluster-Koeffizient

Definition

Definition

Der globale Cluster-Koeffizient C ist definiert als

$$C := \frac{3 \cdot \text{Zahl der (geschlossenen) Dreiecke}}{\text{Zahl der zusammenhängenden Dreiergruppen}}$$

Alternative Definition:

Definition

Der globale Cluster-Koeffizient C ist definiert als

$$C := \frac{6 \cdot \text{Zahl der (geschlossenen) Dreiecke}}{\text{Zahl der Wege der Länge 2}}$$

Beispiel

- Siehe Tafel

Beispiel

- Siehe Tafel

(1, 2, 3)	(4, 2, 1)
(1, 2, 4)	(4, 2, 3)
(1, 2, 5)	(4, 2, 5)
(2, 3, 4)	(4, 3, 2)
(2, 3, 5)	(4, 3, 5)
(2, 3, 6)	(4, 3, 6)
(2, 4, 3)	(4, 6, 3)
(2, 4, 6)	(5, 2, 1)
(2, 5, 3)	(5, 2, 3)
(3, 2, 1)	(5, 2, 4)
(3, 2, 4)	(5, 3, 2)
(3, 2, 5)	(5, 3, 4)
(3, 4, 2)	(5, 3, 6)
(3, 4, 6)	(6, 3, 2)
(3, 5, 2)	(6, 3, 4)
(3, 6, 4)	(6, 3, 5)
	(6, 4, 2)
	(6, 4, 3)

$\Rightarrow$ 34 Wege der Länge 2, 3 Dreiecke

$\Rightarrow C = 18/34$

Bedeutung

- Durchschnittliche Wahrscheinlichkeit, dass zwei Personen mit einem gemeinsamen Freund auch befreundet sind.

$\rightarrow$ C liegt im Intervall $[0, 1]$.

- Ist der Freund meines Freundes mein Freund?

$\rightarrow$ Nicht unbedingt, aber viel wahrscheinlicher als jemand Zufälliges!

- Bei $C = 1$: Alle zusammenhängenden Dreiergruppen bilden eine Clique.
- Frage: Wie hoch ist C in einem Baum?

Diskussion

- Soziale Netzwerke: C oft im zweistelligen Prozent-Bereich.
- Bei technischen Netzwerken ist C häufig niedriger.
- Bsp.: AS-Netzwerk hat nur ca. $C = 0.01$.

Diskussion

- Soziale Netzwerke: C oft im zweistelligen Prozent-Bereich.
- Bei technischen Netzwerken ist C häufig niedriger.
- Bsp.: AS-Netzwerk hat nur ca. $C = 0.01$.

Intuitive Signifikanz der Werte:

- Annahme: Alle Knoten haben etwa Grad c
- Annahme: Meine Freunde wählen ihre Freunde zufällig

$\rightarrow$ Wkt., dass ein solcher Freund auch mein Freund ist, ist etwa c/n

Zwischenfazit Cluster-Koeffizienten

- Cluster-Koeffizient gibt durchschnittliche Dichte von Dreiecken an (global bzw. lokal)
- Maß zur Bestimmung von Netzwerktypen mit bestimmten Eigenschaften
- Auch für gerichtete Graphen möglich bei geeigneter Abwandlung der Definition

Zwischenfazit Cluster-Koeffizienten

- Cluster-Koeffizient gibt durchschnittliche Dichte von Dreiecken an (global bzw. lokal)
- Maß zur Bestimmung von Netzwerktypen mit bestimmten Eigenschaften
- Auch für gerichtete Graphen möglich bei geeigneter Abwandlung der Definition

- Laufzeit der exakten Berechnung des globalen CK mit naivem Algorithmus abhängig von Knotengradverteilung: $O(nd_{max}^2)$
- Bester bekannter exakter Algorithmus: $O(\text{MM}(n))$ mit $\text{MM}(\text{N}) \approx n^{2,373}$
- Approximationsalgorithmus: Deutlich bessere Laufzeit