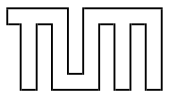
Fortgeschrittene Netzwerk- und Graph-Algorithmen

Dr. Hanjo Täubig

Lehrstuhl für Effiziente Algorithmen
(Prof. Dr. Ernst W. Mayr)
Institut für Informatik
Technische Universität München

Wintersemester 2007/08



Übersicht

- Abgeleitete Kantenzentralitäten
- Vitalität
- Elektrischer Fluss
- Feedback-Zentralitäten

Kantengraph

Definition

Der Kantengraph des Graphen $G = (V, E)$ ist definiert als $G' = (E, K)$, wobei K die Menge der Kanten $e = ((x, y), (y, z))$ mit $(x, y) \in E$ und $(y, z) \in E$ ist.

Zwei Knoten im Kantengraph sind also benachbart, wenn die entsprechenden Kanten im ursprünglichen Graphen einen Knoten gemeinsam haben (im gerichteten Fall als Zielknoten der einen Kante und Startknoten der anderen).

$\Rightarrow$ Wende Knotenzentralität auf den Kantengraph an

Nachteile:

- Größe des Kantengraphen kann quadratisch in der Größe des Graphen sein,
- keine natürliche Interpretation / Generalisierung

Inzidenzgraph

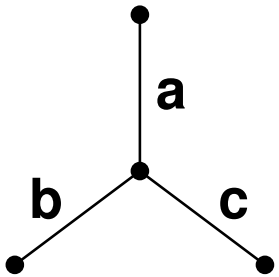
Definition

Der Inzidenzgraph des Graphen $G = (V, E)$ ist definiert als $G'' = (V'', E'')$, wobei $V'' = V \cup E$ und
$E'' = \{(v, e) | \exists w : e = (v, w) \in E\} \cup \{(e, w) | \exists v : e = (v, w) \in E\})$.

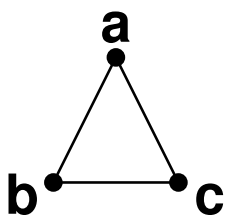
Im Inzidenzgraphen sind also ein 'echter Knoten' und ein 'Kantenknoten' benachbart, wenn der entsprechende Knoten und die entsprechende Kante im ursprünglichen Graphen inzident sind.

$\Rightarrow$ Wende Knotenzentralität auf den Inzidenzgraph an, wobei nur die Pfade zwischen 'echten Knoten' als relevant betrachtet werden

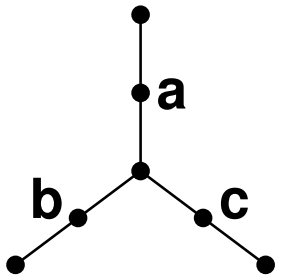
Abgeleitete Kantenzentralitäten



Graph



Kantengraph



Inzidenzgraph

Vitalität

Vitalitätsmaße bewerten die Wichtigkeit eines Knotens oder einer Kante anhand eines Qualitätsverlusts durch das Löschen des Knotens bzw. der Kante.

Definition (Vitalitätsindex)

Sei $\mathcal{G}$ die Menge aller einfachen ungerichteten ungewichteten Graphen $G = (V, E)$ und $f : \mathcal{G} \to \mathbb{R}$ eine reellwertige Funktion auf $G \in \mathcal{G}$.
Dann ist ein Vitalitätsindex $\mathcal{V}(G, x)$ definiert als die Differenz der Werte von f auf G und $G \setminus \{x\}$:

$$\mathcal{V}(G, x) = f(G) - f(G \setminus \{x\})$$

Flow Betweenness Vitality

- ähnlich zu Shortest Path Betweenness
- Motivation: Information in einem Kommunikationsnetzwerk muss sich nicht unbedingt auf kürzesten Pfaden bewegen.
- Maß: Abhängigkeit des maximalen Flusses zwischen zwei Knoten von der Existenz des betrachteten Knotens

Flow Betweenness Vitality

- f_{st}: Maximum-Fluss zwischen Knoten s und t unter Berücksichtigung der Kantenkapazitäten
- $f_{st}(v)$: Fluss zwischen Knoten s und t, der bei Maximum-Fluss von s nach t durch v gehen muss
- $\tilde{f}_{st}(v)$: Maximum-Fluss von s nach t in $G - v$

$$c_{mf}(v) = \sum_{\substack{s,t \in V \\ v \neq s, v \neq t \\ f_{st} > 0}} \frac{f_{st}(v)}{f_{st}} = \sum_{\substack{s,t \in V \\ v \neq s, v \neq t \\ f_{st} > 0}} \frac{f_{st} - \tilde{f}_{st}(v)}{f_{st}}$$

Closeness Vitality

- Wiener Index:

$$I_W(G) = \sum_{v \in V} \sum_{w \in V} d(v, w)$$

- oder in Abhängigkeit der Closeness-Zentralitäten:

$$I_W(G) = \sum_{v \in V} \frac{1}{c_C(v)}$$

- Closeness Vitality für Knoten/Kante x:

$$c_{CV}(x) = I_W(G) - I_W(G - x)$$

- Interpretation: Um wieviel steigen die Gesamtkommunikationskosten, wenn jeder Knoten mit jedem kommuniziert?

Closeness Vitality - Durchschnitt

- Durchschnittliche Distanz:

$$\overline{d}(G) = \frac{I_W(G)}{n(n-1)}$$

- Das Vitalitätsmaß auf der Basis $f(G) = \overline{d}(G)$ misst die durchschnittliche Verschlechterung der Entfernung zwischen zwei Knoten beim Entfernen eines Knotens / einer Kante x.

- Vorsicht! Beim Entfernen eines Artikulationsknotens (cut vertex) oder einer Brücke (cut edge) x ist $c_{CV}(x) = -\infty$.

Shortcut-Werte

- kein echter Vitalitätsindex im Sinne der Definition
- maximale Erhöhung eines Distanzwerts, wenn Kante $e = (v, w)$ entfernt wird

$\Rightarrow$ nur relevant für Knoten, bei denen alle kürzesten Pfade über e laufen

- maximale Erhöhung tritt direkt zwischen Knoten v und w auf
- alternativ: maximale relative Erhöhung
- Berechnung: mit $m = |E|$ Single Source Shortest Path Instanzen (und Vergleich mit der jeweiligen Kante)
- später: mit $n = |V|$ SSSP-Bäumen
- Anwendung auch auf Knotenlöschungen möglich

Stress-Zentralität als Vitalitätsindex

- Stress-Zentralität zählt die Anzahl der kürzesten Pfade, an denen ein Knoten oder eine Kante beteiligt ist

$\Rightarrow$ kann als Vitalitätsmaß betrachtet werden

- Aber: Anzahl der kürzesten Pfade kann sich durch Löschung erhöhen (wenn sich die Distanz zwischen zwei Knoten erhöht)

$\Rightarrow$ Längere kürzeste Pfade müssen ignoriert werden

Stress-Zentralität als Vitalitätsindex

- $f(G \setminus \{v\})$ muss ersetzt werden durch

$$\sum_{s \in V} \sum_{t \in v} \sigma_{st}[d_G(s,t) = d_{G \setminus \{v\}}](s,t)$$

- Ausdruck in Klammern ist 1 (wahr) oder 0 (falsch)
- $f(G \setminus \{e\})$ analog:

$$\sum_{s \in V} \sum_{t \in v} \sigma_{st}(e)[d_G(s,t) = d_{G \setminus \{e\}}](s,t)$$

- kein echter Vitalitätsindex im Sinne der Definition, weil der Index-Wert nach der Löschung von der Distanz vor der Löschung abhängt

Elektrische Netzwerke

Elektrisches Netzwerk

- einfacher ungerichteter zusammenhängender Graph $G = (V, E)$
- Leitwertfunktion $c : E \to \mathbb{R}$
- Einspeisung $b : V \to \mathbb{R}$
  - positive Werte: eingehender Strom
  - negative Werte: ausgehender Strom
- Bedingung

$$\sum_{v \in V} b(v) = 0$$

- Kanten erhalten beliebige Orientierung $\vec{E}$

Elektrischer Fluss

Definition (Kirchhoffsche Gesetze)

Eine Funktion $x: E \to \mathbb{R}$ heißt (elektrischer) Strom falls

$$\forall v \in V: \quad \sum_{(v,w)\in\vec{E}} x(v,w) - \sum_{(w,v)\in\vec{E}} x(w,v) = b(v)$$

und

$$\sum_{e\in C} x(\vec{e}) = 0$$

für alle Zyklen (Kreise) C des Graphen G
Negative Werte bedeuten Fluss entgegengesetzt zur Kantenrichtung.

Elektrisches Potential

Definition

Eine Funktion $p: \ V \to \mathbb{R}$ heißt (elektrisches) Potential falls

$$\forall (v,w) \in \vec{E}: \quad p(v) - p(w) = \frac{x(v,w)}{c(v,w)}$$

- elektrisches Netzwerk $N = (G, c)$ hat einen eindeutig bestimmten Strom für jede Einspeisung b,
- ebenso Potential (bis auf einen additiven Term), also eigentlich Potentialdifferenzen (Spannung)

Laplacian Matrix

- $L = L(N)$
- $$L_{vw} = \begin{cases} \sum_{e \ni v} c(e) & \text{if } v = w \\ -c(e) & \text{if } e = \{v, w\} \\ 0 & \text{sonst} \end{cases}$$
- Für gegebenes Netzwerk $N = (G, c)$ und Einspeisung b kann man ein Potential durch Lösen der Gleichung $Lp = b$ finden.
- unit s-t-supply b_{st}:
  $b_{st}(s) = 1,\ b_{st}(t) = -1,\ \forall v \in V \setminus \{s, t\}$

Current-Flow Betweenness Centrality

- Durchsatz eines Knotens v bezüglich einer s-t-Einheitsspeisung b_{st}:

$$\tau_{st}(v) = \frac{1}{2}\left(-|b_{st}(v)| + \sum_{e \ni v} |x(\vec{e})|\right)$$

- $-|b_{st}(v)|$: Durchsatz für Anschlussknoten auf Null
- $\frac{1}{2}$ weil ein- und ausgehender Strom aufsummiert wird
- Current Flow Betweenness:

$$c_{CB}(v) = \frac{1}{(n-1)(n-2)} \sum_{s,t \in V} \tau_{st}(v)$$

Current-Flow Closeness Centrality

- Current-Flow Closeness:

$$c_{CC}(v) = \frac{n-1}{\sum_{t \neq v} p_{vt}(v) - p_{vt}(t)}$$

- Brandes / Fleischer:
  Current-Flow Closeness = Informationszentralität
- Informationszentralität:

$$c_I(v)^{-1} = nM_{vv} + \text{trace}(M) - \frac{2}{n},$$

wobei $M = (L + U)^{-1}$, L ist die Laplacian und U ist gleichgroße Matrix mit Einsen
(trace(M): Summe der Diagonalelemente)

Zufallsprozesse

- Was tun, wenn man kürzeste Pfade nicht berechnen kann? (z.B. weil man nicht das ganze Netzwerk kennt, sondern nur einen lokalen Ausschnitt)

$\Rightarrow$ zufällig einen Nachbarn des aktuellen Knotens explorieren (Random Walk)

- Beispiel: Weitergabe von Banknoten

Random Walks und Gradzentralität

Satz

In ungerichteten Graphen sind die Wahrscheinlichkeiten der stationären Verteilung bei einem kanonischen Random Walk proportionl zum Knotengrad.

$$p_{ij} = \frac{a_{ij}}{\deg(i)} \quad \Rightarrow \quad \pi_i = \frac{\deg(i)}{\sum_{v \in V} \deg(v)} = \frac{\deg(i)}{2m}$$

Random Walks und Gradzentralität

Beweis.

$$
\begin{aligned}
(\pi P)_j &= \sum_{i \in V} \pi_i p_{ij} = \frac{\sum_{i \in V} \deg(i) p_{ij}}{\sum_{v \in V} \deg(v)} \\
&= \frac{\sum_{i \in V} a_{ij}}{\sum_{v \in V} \deg(v)} = \frac{d(j)}{\sum_{v \in V} \deg(v)} = \pi_j
\end{aligned}
$$

□

π bzw. π_i: Wahrscheinlichkeit(en) der stationären Verteilung
P bzw. p_{ij}: Übergangswahrscheinlichkeit(en) von Knoten i nach j
a_{ij}: Eintrag (i, j) der Adjazenzmatrix

Random Walk Betweenness Centrality

- Knoten s will Knoten t eine Nachricht schicken, kennt aber nicht den kürzesten Weg

$\Rightarrow$ Random Walk

- Falls die Nachricht bei t ankommt, wird sie absorbiert
- M bzw. m_{ij}: Wahrscheinlichkeit, dass Knoten j die Nachricht an Knoten i weiterschickt

$$m_{ij} = \begin{cases} \frac{a_{ij}}{\deg(j)} & \text{falls } j \neq t \\ 0 & \text{sonst} \end{cases}$$

Random Walk Betweenness Centrality

- D: Grad-Matrix

$$d_{ij} = \begin{cases} \deg(i) & \text{falls } i = j \\ 0 & \text{sonst} \end{cases}$$

- D^{-1}: Inverse von D mit Reziproken Knotengraden auf der Hauptdiagonale (sonst Null)
- Durch das Absorptionsverhalten ist die folgende Darstellung nicht ganz korrekt

$$M = A \cdot D^{-1}$$

$\Rightarrow$ Löschen von Zeile t und Spalte t:

$$M_t = A_t \cdot D_t^{-1}$$

(Index t gibt die gelöschte Zeile/Spalte an)

Random Walk Betweenness Centrality

- Alle möglichen Pfade betrachten, zusammen mit der jeweiligen Wahrscheinlichkeit, dass dieser Pfad ausgewählt wird
- Wieviele verschiedene Pfade der Länge r von Knoten i nach Knoten j gibt es?

$$(A^r)_{ij}$$

- Wahrscheinlichkeit, dass ein in s gestarteter Random Walk sich nach r Schritten bei Knoten j befindet:

$$(M_t^r)_{js}$$

- Wahrscheinlichkeit, dass sich in Schritt $r+1$ der Random Walk zu Knoten i bewegt:

$$\left(M_t^{r+1}\right)_{js} = m_{ij}^{-1} \left(M_t^r\right)_{js}$$

Random Walk Betweenness Centrality

- Wahrscheinlichkeit, dass Knoten j eine in s gestartete Nachricht zu Knoten i schickt, summiert über alle Pfade der Länge 0 bis ∞:

$$\sum_{r=0}^{\infty} m_{ij}^{-1} \left(M_t^r\right)_{js} = m_{ij}^{-1} \left[\left(I_{n-1} - M_t\right)^{-1}\right]_{js}$$

  I_{n-1} ist die Identitätsmatrix der Dimension $n-1$.
- Alle Einträge in M_t^r sind Wert zwischen 0 und 1, die Summe ist also konvergent
  (werden wir später bei den Feedback-Zentralitäten sehen).

Random Walk Betweenness Centrality

- Sei $\mathbf{s}$ ein Vektor der Dimension $n-1$, der 1 ist bei Knoten s, und sonst 0.
- Dann kann man die letzte Gleichung wie folgt umschreiben:

$$\begin{aligned} \mathbf{v^{st}} &= D_t^{-1} \cdot (I_{n-1} - M_t)^{-1} \cdot \mathbf{s} \\ &= (D_t - A_t)^{-1} \cdot \mathbf{s} \end{aligned}$$

- Vektor $\mathbf{v^{st}}$ beschreibt die Wahrscheinlichkeit, die Nachricht bei Knoten i zu finden, während sie auf dem Weg von s nach t ist.

Random Walk Betweenness Centrality

- Einige Random Walks haben redundante Teile, weil sie zu einem Knoten zurückkehren, bei dem sie zuvor schon gewesen sind.
- Netzwerk ist ungerichtet

$\Rightarrow$ Kreise können in beiden Richtungen durchlaufen werden und löschen sich aus

- $\mathbf{v}^{\mathbf{st}}$ ignoriert diese Kreise

Random Walk Betweenness Centrality

⇒ Analogie zu Strom-Fluss in elektrischen Netzwerken

- el. Netzwerk $N = (G, c)$ mit einheitlichen Kantengewichten $c(e) = 1 \quad (\forall e \in E)$
- Laplacian $L(N) = D - A$
  (D: Grad-Matrix, A: Adjazenzmatrix)

⇒ Ein Potential p_{st} in N für ein 'unit s-t-supply' b_{st} ist eine Lösung für das System $Lp_{st} = b_{st}$.

- Matrix L hat nicht vollen Rang ⇒ ein Potential festlegen
  (Knoten v, Potentiale sind eindeutig bis auf additiven Term)

⇒ Matrizen L_v, D_v und A_v
(Zeilen und Spalten zu v gelöscht)

- L_v hat vollen Rang ⇒ invertierbar

⇒ $p_{st} = L_v^{-1} b_{st} = (D_v - A_v)^{-1} b_{st}$

- entspricht $\mathbf{v^{st}} = (D_t - A_t)^{-1} \cdot \mathbf{s}$

⇒ Analogie zwischen RW Betw. und Current-Flow Betw.:
$c_{RW\ B}(v) = c_{C\ B}(v)$

Random Walk Closeness Centrality

- gleicher Ansatz liefert Random Walk Closeness
- Mean First Passage Time (MFPT) m_{st}:
  erwartete Anzahl Knoten, die vom Start in s bis zur (ersten) Ankunft in t besucht werden

$$m_{st} = \sum_{n=1}^{\infty} n \cdot f_{st}^{(n)}$$

  $f_{st}^{(n)}$: Wahrscheinlichkeit, dass man nach genau n Schritten erstmals bei t ankommt
- $M = (I - EZ_{dg})D \ldots$

Markov-Zentralität

- Markov-Zentralität:

$$c_M(v) = \frac{n}{\sum_{s \in V} m_{sv}}$$

- sinnvoll gerichtete und ungerichtete Graphen
  in gerichteten: durchschnittliche MFPT für Random Walks, die in v beginnen bzw. enden
- erwartete Anzahl von Schritten von v zu den anderen Knoten (oder von den anderen zu v)

$\Rightarrow$ eine Art durchschnittliche Random Walk Closeness

Rückkopplungen

- Ein Knoten ist umso zentraler, je zentraler seine Nachbarn sind. (Aber ein Knoten ist ja auch ein Nachbar seiner Nachbarn...)

- Notation: Vektoren statt Funktionen (für lineare Gleichungssysteme)

Der Status-Index von Katz

- Beispiel: indirekte Wahl (z.B. k und l wählen i, aber alle anderen wählen nur k oder l und damit indirekt i)

$\Rightarrow$ Zähle auch die indirekten Stimmen, aber unter Berücksichtigung eines Dämpfungsfaktors $\alpha > 0$ (je länger der Pfad, desto geringer der Einfluss)

- einfacher ungewichteter gerichteter Graph mit Adjazenzmatrix A
- $(A^k)_{ji}$: Anzahl Pfade von j nach i der Länge k
- Katz' Status-Index: (bei Konvergenz)

$$c_K(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji} \quad \text{bzw.} \quad \mathbf{c_K} = \sum_{k=1}^{\infty} \alpha^k (A^T)^k \mathbf{1}_n$$

Katz' Status - Konvergenz

Um Konvergenz zu garantieren, muss α beschränkt werden.

Satz

Sei A die Adjazenzmatrix von G, $\alpha > 0$ und λ_1 der größte Eigenwert von A. Dann gilt:

$$\lambda_1 < \frac{1}{\alpha} \quad \Leftrightarrow \quad \sum_{k=1}^{\infty} \alpha^k A^k \text{ konvergiert}$$

Katz' Status - Geschlossene Form

- Bei Konvergenz erhält man die geschlossene Form

$$\mathbf{c_K} = \sum_{k=1}^{\infty} \alpha^k (A^T)^k \mathbf{1}_n = \left((I - \alpha A^T)^{-1} \right) \mathbf{1}_n$$

bzw.

$$(I - \alpha A^T)\mathbf{c_K} = \mathbf{1}_n$$

- inhomogenes lineares Gleichungssystem,
Rückkopplung: $c_K(i)$ hängt von den anderen $c_K(j)$ mit $j \neq i$ ab