

Web Algorithmen Clustering

Dr. Michael Brinkmeier

Technische Universität Ilmenau
Fakultät Informatik und Automatisierung
Fachgebiet Automaten und Formale Sprachen

8.7.2005

Teil I

λ -Mengen

Flake's Communities (Fortsetzung)

Definition (Flake et al. 2000, Zur Erinnerung)

Sei $G = (V, E)$ ein ungerichteter, gewichteter Graph mit Kantengewichten $w(u, v)$. Eine **Community** in G ist eine Teilmenge $C \subset V$ von Knoten, so dass

$$w(v, C) \geq w(v, V \setminus C).$$

Borgatti, Everett und Shirey benutzten diese Definition bereits 1990 unter dem Namen **α -Menge**.

Die Berechnung erfolgt über minimale S - t -Schnitte, wobei

- S die Saat, eine Menge von ausgezeichneten Knoten, und
- t eine künstliche Senke ist.

Flake's Communities (Fortsetzung)

Vorschlag

Statt einen minimalen S - t -Schnitt für eine künstliche Senke zu berechnen, berechne

$$\min \{ \lambda_G(S, v) \mid v \in V \setminus S \}$$

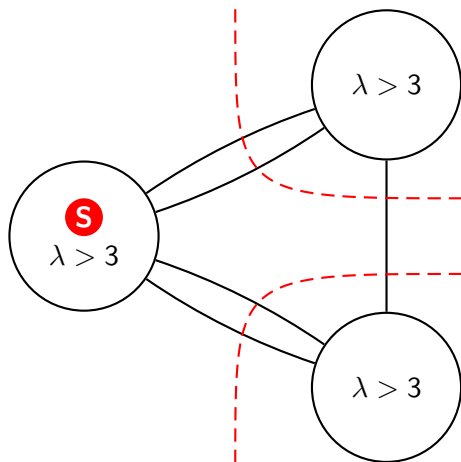
und wähle einen minimalen S -Schnitt als Community.

Definition (S - T -Schnitt)

Sei $G = (V, E)$ ein ungerichteter, gewichteter Graph und $S, T \subset V$. Ein **S - T -Schnitt** von G ist ein Schnitt $S' \subsetneq V$ mit $S \subseteq S'$ und $T \not\subseteq S'$.

Das Gewicht eines minimalen S - T -Schnittes S' bezeichnen wir mit $\lambda_G(S, T)$.

Flake's Communities (Fortsetzung)



Fazit: Diese Communities sind wieder nicht eindeutig.

Flake's Communities (Fortsetzung)

Um Eindeutigkeit zu erzwingen, könnte man den Schnitt aller minimalen S -Schnitte als Community von S wählen.

Dies ergibt das folgende vorläufige Definition:

Definition (vorläufig)

Sei $\lambda := \min \{ \lambda_G(S, v) \mid v \notin S \}$. Dann sei

$$\text{Community}(S) := \{ u \mid \lambda_G(S, u) > \lambda \}$$

λ -Mengen

Definition (λ -Mengen)

Sei $G = (V, E)$ ein gewichteter, ungerichteter Graph und $k \in \mathbb{R}$. Eine **k - λ -Menge** von G ist eine Menge X von Knoten, so dass für alle $u, v \in X$ und $w \notin X$ folgendes gilt:

$$\lambda_G(u, v) > k \text{ und } \lambda_G(u, w) \leq k.$$

Fakt

Sei $F \subseteq E$ die Menge aller Kanten, die in Schnitten mit einem Gewicht $\leq k$ liegen. Dann sind die k - λ -Mengen von G die zusammenhängenden Komponenten von $G' := (V, E \setminus F)$.

Insbesondere bilden die k - λ -Mengen eine Partition der Knotenmenge V .

λ -Mengen

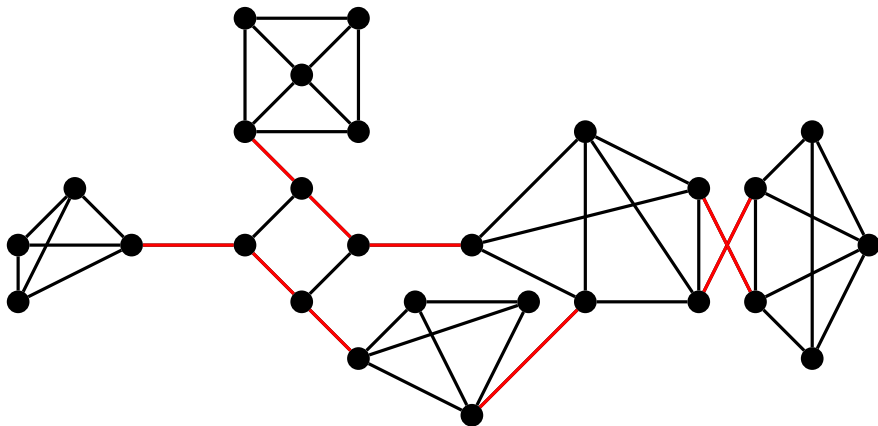
Folgerung

Sei $k' > k$. Dann gibt es für jede k' - λ -Menge X eine k - λ -Menge Y , so dass $X \subseteq Y$.

D.h. die Partitionen werden mit steigendem k *feiner*.

Die λ -Mengen können mittels sogen. **Schnitt-Bäume** (Gomory, Hu 1961) berechnet werden.

λ -Mengen - ein Beispiel

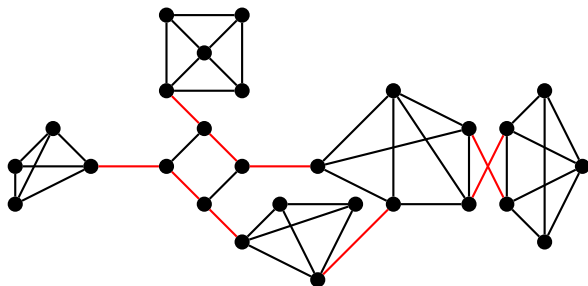


Ein neuer Community-Begriff ?

Bei einer k - λ -Menge X kann die interne Bedingung

$$\lambda_G(u, v) > k$$

auch über Wege erreicht werden, die **außerhalb** der Menge X liegen.



Ein neuer Community-Begriff ?

Um dies zu vermeiden könnte man die Definition von λ -Mengen strikter gestalten, indem man den inneren Zusammenhang nur über die Menge selbst realisiert.

Definition (Strikte λ -Mengen)

Eine **strikte k - λ -Menge** X in einem ungerichteten, gewichteten Graphen $G = (V, E)$ ist eine Menge $X \subseteq V$, so dass für $u, v \in X$ und $w \notin X$ Folgendes erfüllt ist:

$$\lambda_{G[X]}(u, v) > k \text{ und } \lambda_G(u, w) \leq k.$$

Dabei ist $G[X]$ der von der Knotenmenge X induzierte Untergraph.

Es bleibt die Frage, ob diese Bedingung erfüllt werden kann und wie die Mengen berechnet werden können.

Teil II

Communities und k -Komponenten

Kantenzusammenhang (Wiederholung)

Definition (Schnitt und Kantenzusammenhang)

Sei $G = (V, E)$ ein ungerichteter, gewichteter Graph. Ein **(Kanten-)Schnitt** von G ist eine Menge $S \subsetneq V$ von Knoten.

Ein Schnitt S heisst **minimal**, wenn er minimales Gewicht $w(S)$ unter allen Schnitten von G hat, d.h.

$$w(S) = \min \{w(U) \mid U \subsetneq V\}.$$

Das Gewicht eines minimalen Schnittes S von G nennen wir **Kantenzusammenhang** und bezeichnen es mit $\lambda(G)$.

Fakt

Seien H_1 und H_2 Teilgraphen von G mit $H_1 \cap H_2 \neq \emptyset$, dann gilt

$$\lambda(H_1 \cup H_2) \geq \min(\lambda(H_1), \lambda(H_2)).$$

Communities

Eine andere Möglichkeit den Zusammenhang **innerhalb** der Community zu garantieren ergibt sich über die folgende Definition.

Definition (Communities)

Sei $G = (V, E)$ ein ungerichteter, gewichteter Graph. Eine **Community** eines Teilgraphen H von G ist ein Teilgraph C , der die folgenden Bedingungen erfüllt:

- 1 H ist in C enthalten.
- 2 C hat maximalen Kantenzusammenhang unter allen Teilgraphen von G , die H enthalten, d.h. für alle $D \subseteq G$ mit $H \subseteq D$ gilt:

$$\lambda(D) \leq \lambda(C).$$

- 3 C ist maximal unter allen Teilgraphen mit maximalen Kantenzusammenhang, die H enthalten, d.h. für alle $D \subseteq G$ mit $H \subseteq D$ und $\lambda(D) = \lambda(C)$ gilt $D \subseteq C$.

Existenz und Eindeutigkeit der Communities

Lemma (Existenz und Eindeutigkeit der Communities)

Für jeden Teilgraphen $H \subseteq G$ existiert eine eindeutig bestimmte Community $\text{Comm}_G(H)$.

Beweis.

Sei \mathcal{H} die Menge aller Teilgraphen D von G mit $H \subseteq D$ und maximalem $\lambda(D)$ unter allen Teilgraphen, die H enthalten. Dann gilt

$$\lambda\left(\bigcup_{D \in \mathcal{H}} D\right) \geq \min\{\lambda(D) \mid D \in \mathcal{H}\}$$

und wegen der Maximalität von $\lambda(D)$ sogar die Gleichheit. □

Definition (Stärke von H in G)

$$\text{str}_G(H) := \lambda(\text{Comm}_G(H))$$

Fakt

Sei H ein Teilgraph von $G = (V, E)$.

- 1 Jede Community ist ein induzierter Untergraph, d.h. es gibt eine Knotenmenge U , so dass

$$\text{Comm}_G(H) = G[U].$$

- 2 Jede Community wird von einem Knoten oder einer Kante *erzeugt*, d.h. es gibt einen Knoten oder eine Kante $x \in H$, so dass

$$\text{Comm}_G(H) = \text{Comm}_G(x).$$

Communities und minimale Schnitte

Lemma

Sei S ein minimaler Schnitt von $G = (V, E)$ und H ein Teilgraph von G .
Dann gilt

$$\text{str}_G(H) > \lambda(G) \Leftrightarrow \text{Comm}_G(H) \subseteq G[S] \text{ oder } \text{Comm}_G(H) \subseteq G[\bar{S}]$$

und

$$\text{str}_G(H) = \lambda(G) \Leftrightarrow \text{Comm}_G(H) = G.$$

Anders ausgedrückt: Die Community eines Teilgraphen liegt entweder vollständig in einer Hälfte eines minimalen Schnittes, oder ist der gesamte Graph.

Satz

Seien H_1 und H_2 zwei Teilgraphen von $G = (V, E)$. Dann tritt genau einer von folgenden vier Fällen ein:

- ① $\text{Comm}_G(H_1) = \text{Comm}_G(H_2)$
- ② $\text{Comm}_G(H_1) \subsetneq \text{Comm}_G(H_2)$
- ③ $\text{Comm}_G(H_2) \subsetneq \text{Comm}_G(H_1)$
- ④ $\text{Comm}_G(H_1) \cap \text{Comm}_G(H_2) = \emptyset$

Anders formuliert: Die Communities von allen Teilgraphen in G bilden einen Baum, wobei $\text{Comm}_G(H_1)$ genau dann ein Kind von $\text{Comm}_G(H_2)$ ist, wenn

- $\text{Comm}_G(H_1) \subsetneq \text{Comm}_G(H_2)$ und
- keine Community dazwischen existiert.

Minimale-Schnitt-Bäume

Definition

Ein **Minimaler-Schnitt-Baum** T eines ungerichteten, gewichteten Graphen $G = (V, E)$ ist ein binärer Baum mit Wurzel, so dass

- jedes Blatt von T genau einem Knoten von G entspricht und umgekehrt.
- zu jedem inneren Knoten x des Baumes T die Menge

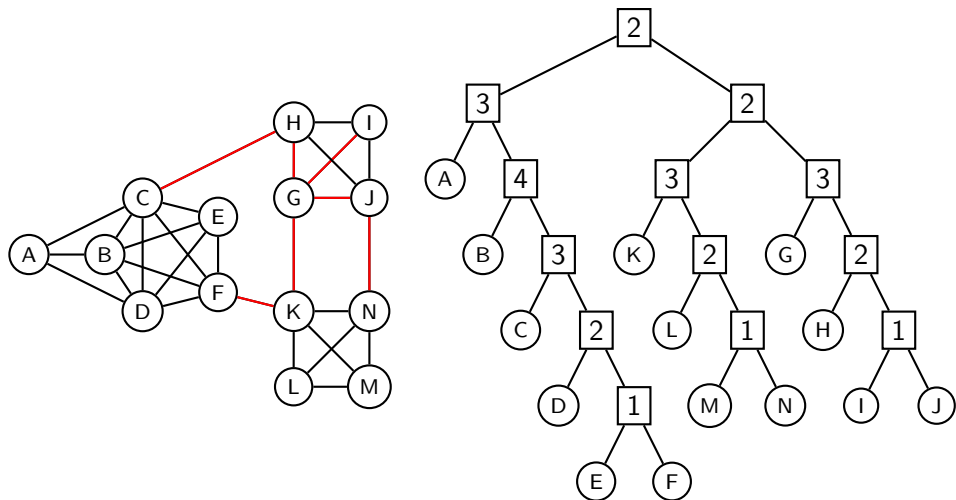
$$V(x) = \{v \in V \mid v \text{ liegt im Teilbaum mit Wurzel } x\}$$

ein minimaler Schnitt des vom Vorgängers induzierten Graphen ist, d.h. ist y der Vorgänger von x in T , so gilt

$$V(x) \text{ ist ein minimaler Schnitt von } G[V(y)].$$

Die Wurzel und die inneren Knoten von T erhalten das Gewicht des durch die Kinder induzierten Schnittes als Gewicht $\lambda(x)$.

Minimale-Schnitt-Bäume – ein Beispiel



Berechnung der Communities

Ist ein Minimaler-Schnitt-Baum T von G gegeben, lässt sich die Community eines Teilgraphen H leicht finden:

- 1 Markiere alle Blätter, die Knoten in H entsprechen.
- 2 Finde die Wurzel r des kleinsten Teilbaumes, der alle markierten Blätter enthält.
- 3 Gehe von r zur Wurzel von T und suche den letzten Knoten x auf diesem Weg mit möglichst hohem Gewicht $\lambda(x)$.

Dann ist $G[V(x)]$ die Community von H .

Die Berechnung von Communities

Hat man einen Minimalen-Schnitt-Baum T , so kann man alle Knoten, die Communities darstellen, markieren und die übrigen Knoten entfernen. Dadurch erhält man den **Community-Baum** $CT(G)$.

Satz

Sei $G = (V, E)$ ein ungerichteter, gewichteter Graph. Existiert ein Algorithmus zur Berechnung eines minimalen Schnittes eines unger., gew. Graphen mit Laufzeit $O(g(|V|, |E|))$, dann kann $CT(G)$ in Zeit $O(|V|g(|V|, |E|))$ berechnet werden.

Für die Berechnung von minimalen Schnitten existieren mehrere Algorithmen:

- 1 Nagamochi & Ibaraki (1992): $O(|V||E| + |V|^2 \log |V|)$
- 2 Hao & Orlin (1994): $O(|V||E| \log \frac{|V|^2}{|E|})$
- 3 Brinkmeier (2005): $O(\delta(G)|V|^2)$ (ganzzahlige Kantengewichte)

Strikte λ -Mengen und Communities

Satz

In jeden ungerichteten, gewichteten Graphen $G = (V, E)$ sind die strikten λ -Mengen genau die Communities.

Beweis

Zuerst zeigen wir, dass jede Community C eine strikte λ -Menge ist.

Offensichtlich gilt $\lambda_C(u, v) \geq \lambda(C)$ für $u, v \in C$.

C wird durch einen Teilbaum mit Wurzel r in einem

Minimalen-Schnitt-Baum repräsentiert. Dabei haben alle Vorgänger einen

Kantenzusammenhang echt kleiner $\lambda(C)$. Damit ist jeder Knoten $w \notin C$ ein Blatt in einem anderen Teilbaum, der von einem Vorgänger p abzweigt.

Somit gilt für $u \in C$ und $w \notin C$

$$\lambda_G(u, w) \leq \lambda(p) < \lambda(C).$$

Strikte λ -Mengen und Communities

Beweis.

Es bleibt zu zeigen, dass jede strikte λ -Menge eine Community ist.

Sei X eine strikte k - λ -Menge. Damit gilt $\lambda(X) > k$. Sei nun $C = \text{Comm}_G(X)$, was offensichtlich $\lambda(C) \geq \lambda(X) > k$ zur Folge hat.

Nehmen wir nun an, dass ein $w \in C \setminus X$ existiert. Dann gilt einerseits

$$\lambda_C(u, w) > k \text{ f\u00fcr jedes } u \in C,$$

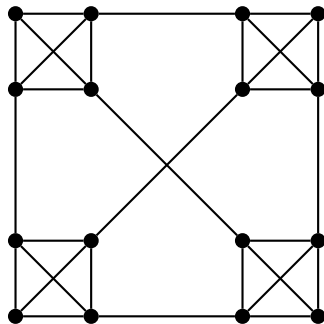
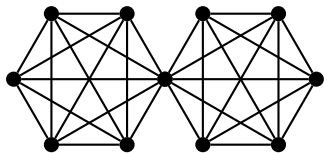
und andererseits

$$\lambda_X(u, w) \leq k \text{ f\u00fcr } u \in X \subseteq C.$$

Da das zu einem Widerspruch f\u00fchrt, kann kein solches w existieren und somit muss $X = C$ gelten. □

Probleme

Die folgenden Strukturen bilden eine einzige Community.



Idee

Statt der maximalen Menge mit maximalem Kantenzusammenhang, suche minimale Mengen mit maximalem Kantenzusammenhang.

Kleinste Community

Problem (MIN-COMMUNITY)

Gegeben: Ein Graph $G = (V, E)$ und $k, \lambda \in \mathbb{N}$

Gesucht: Ein Teilgraph $D \subseteq G$, so dass

- D höchstens k Knoten besitzt,
- $\lambda(D) \geq \lambda$.

Satz

MIN-COMMUNITY ist NP-vollständig.

Kleinste Community

Beweis.

Da der Kantenzusammenhang in polynomieller Zeit berechnet werden kann, gilt offensichtlich $\text{MIN-COMMUNITY} \in NP$.

Um die Vollständigkeit zu zeigen reduzieren wir CLIQUE auf MIN-COMMUNITY .

Wir betrachten nur ungewichtete Graphen. Es ist leicht einzusehen, dass für einen solchen Graphen G mit k Knoten gilt

$$\lambda(G) = k - 1 \Leftrightarrow G \text{ ist vollständig.}$$

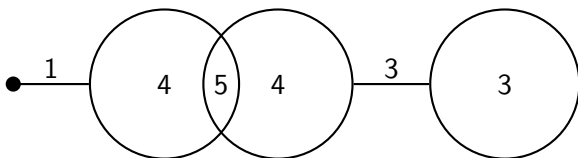
Ausserdem gilt stets $\lambda(G) \leq k - 1$.

Damit gilt

$$(G, k) \in \text{CLIQUE} \Leftrightarrow (G, k, k - 1) \in \text{MIN-COMMUNITY.}$$



Weitere Probleme



Vielen Dank!