

Efficient Detection of Network Motifs

Seminar

"Aktuelle Themen der Bioinformatik"

Referent: Alexander Grimme

Quelle:

„Efficient Detection of Network Motifs“
Sebastian Wernicke

I. Einleitung

- **Grundidee**

- Die Evolution erhält Module (Netzwerk Motive), die eine bestimmte Funktionalität aufweisen

- **Netzwerk Motive**

- kleine, zusammenhängende Teilnetzwerke
- treten in einer signifikant höheren Frequenz auf, als in einem Zufallsnetzwerk

I. Einleitung

- Konzept zur Erforschung struktureller Zusammenhänge in biologischen Netzwerken
- Zählen der einzelnen Netzwerk Motive und vergleiche deren Häufigkeit mit dem Zufallsnetzwerk

Kritik

- Globale Netzwerkeigenschaften beeinflussen lokale Netzwerkeigenschaften

I. Einleitung

- Ziel:
 - Verbesserung der bekannten Algorithmen zur Analyse von Netzwerk Motiven, um komplexere Netzwerke untersuchen zu können und so neue Erkenntnisse zu erlangen.

II. Notation

1. Graphen:

Sei $G = (V, E)$ ein Graph

$$n := |V|$$

Seien alle Knoten $v \in V$ eindeutig nummeriert mit $1, 2, \dots, n$

Alle Kanten werden ungerichtet notiert, also $\{u, v\}$

II. Notation

2. Offene Nachbarschaft $N(V')$:

Menge von allen Knoten $V \setminus V'$, die adjazent zu mindestens einem Knoten aus V' sind.

3. Exklusive Nachbarschaft $N_{\text{excl}}(v, V')$:

Für einen Knoten $v \in V \setminus V'$ ist seine exklusive Nachbarschaft bezüglich V' die Menge aller benachbarten Knoten, die nicht zu $V' \cup N(V')$ gehören.

II. Notation

4. Subgraph-Klassen:

Ein zusammenhängender Subgraph, induziert durch eine Menge M von Knoten, heißt Subgraph der Größe k , genau dann, wenn $|M| = k$.

Für ein festes k , wird die Menge G aller Subgraphen der Größe k in Teilmengen $S_k^i(G)$ unterteilt. Dabei sind zwei Subgraphen der Größe k genau dann in der selben Subgraph-Klasse, wenn sie isomorph (d.h. topologisch äquivalent) sind.

II. Notation

5. Konzentration von Subgraph-Klassen:

$$C_k^i(G) := \frac{|S_k^i(G)|}{\sum_j |S_k^i(G)|}$$

6. Ein Schätzer für $C_k^i(G)$:

Sei G ein Graph und R eine Menge von zufällig erzeugten Subgraphen der Größe k . Dann ist die

Abbildung $\hat{C}_k^i : (R, G) \rightarrow [0, 1]$ ein Schätzer für $C_k^i(G)$

II. Notation

6. Baised/Unbaised:

$\hat{C}_k^i(R, G)$ heißt *unbaised*, wenn gilt:

$$\hat{C}_k^i(R, G) = C_k^i \quad (\text{bezl. des Algorithmus } A)$$

sonst heißt der Schätzer baised.

III. Verfahren

1. Finde alle vorkommenden Subgraphen und deren Häufigkeit
2. Bilde Subgraph-Klassen, so dass alle topologisch gleichen Subgraphen in der selben Subgraph-Klasse liegen
3. Bestimme welche Subgraph-Klassen besonders häufig auftreten im Vergleich zu einem bestimmten Zufallsgraph Modell

III. Subgraph Sampling (ESA)

Eingabe: Graph $G(V, E)$ und Integer $2 \leq k \leq |V|$

Ausgabe: Menge zufällig erzeugter Subgraphen der Größe k

01 $\{u, v\} \leftarrow$ zufällige Kante aus E

02 $V' \leftarrow \{u, v\}$

03 while $|V'| \neq k$ do

04 $\{u, v\} \leftarrow$ zufällige Kante zwischen V' und $N(V')$

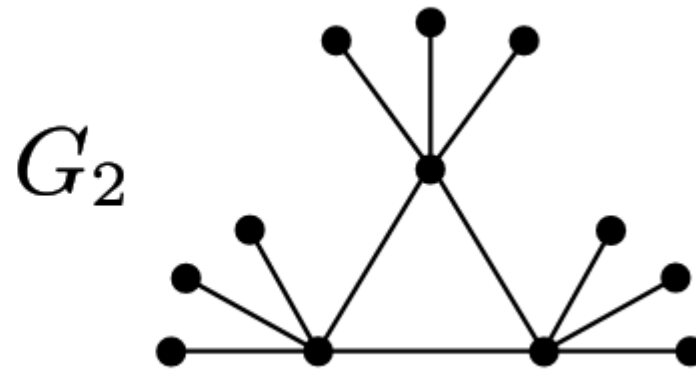
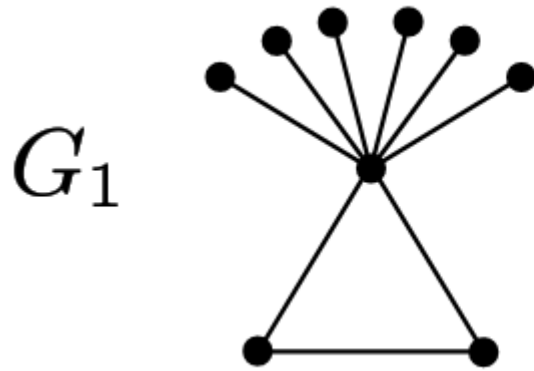
05 $V' \leftarrow V' \cup \{u, v\}$

06 return V'

III. Analyse von ESA

- ESA erzeugt bestimmte Subgraphen häufiger als andere (oversampling)
- Keine Aussage welcher Anteil von Subgraphen zufällig erzeugt wurde
- ESA kann den gleichen Subgraph mehrmals erzeugen, so dass Rechenleistung verbraucht wird, ohne neue Informationen zu erhalten

III. Beispiel: Oversampling



III. Subgraph Enumeration

- Benutze einen Algorithmus, der alle Subgraphen der Größe k aufzählt.
- Erweitere diesen Algorithmus so, dass er zufällig Subgraphen überspringt und nicht berechnet.

=> unbiased Sampling Algorithmus

III. Subgraph Enumeration (ESU)

Eingabe : $G=(V, E)$ und $1 \leq k \leq |V|$

Ausgabe : Alle Subgraphen der Größe k

01 *for alle Knoten* $v \in V$ *do*

02 $V_{\text{Extension}} \leftarrow \{u \in N(\{v\}) : u > v\}$

03 *ExtendSubgraph*($\{v\}, V_{\text{Extension}}, v$)

04 *return*

E1 *if* $|V_{\text{Subgraph}}| = k$ *then* *Ausgabe* $G[V_{\text{Subgraph}}]$ *und* *return*

E2 *while* $V_{\text{Extension}} \neq \emptyset$ *do*

E3 *Entferne einen zufälligen Knoten* w *aus* $V_{\text{Extension}}$

E4 $V'_{\text{Extension}} \leftarrow V_{\text{Extension}} \cup \{u \in N_{\text{excl}}(w, V_{\text{Subgraph}}) : u > v\}$

E5 *ExtendSubgraph*($V_{\text{Subgraph}} \cup \{w\}, V'_{\text{Extension}}, v$)

E6 *return*

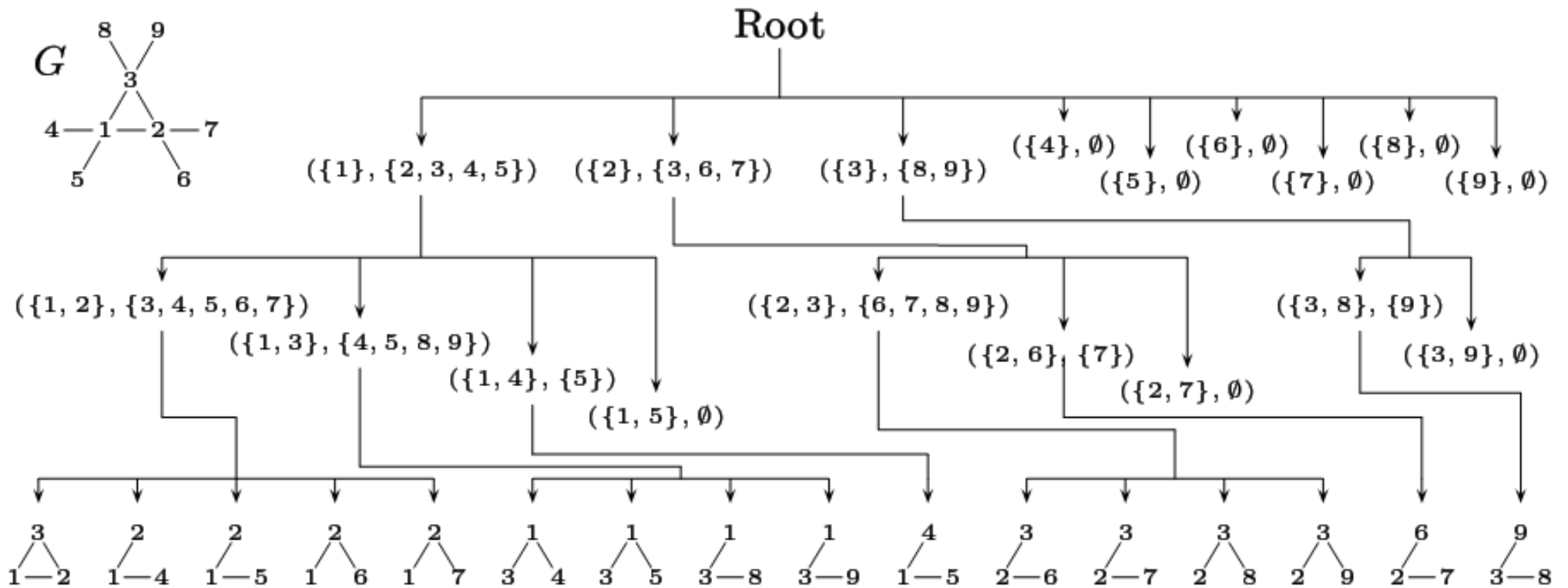
III. ESU-Baum

Definition 1:

Mit einem Aufruf von ESU (G,k) entsteht ein Baum von rekursiven Funktionsaufrufen. Dieser Baum heißt ESU-Baum. Die Wurzel repräsentiert den Aufruf des Algorithmus (*Tiefe Null*). Jeder rekursive Funktionsaufruf wird durch eine Kante vom aufrufenden Knoten zum aufgerufenen Knoten dargestellt. Ein aufgerufener Knoten wird mit $(V_{\text{Subgraph}}, V_{\text{Extension}})$ beschriftet und hat eine Tiefe von $|V_{\text{Subgraph}}|$

III. ESU-Baum

Beispiel



III. ESU-Baum

Notation

Für einem Knoten w im ESU-Baum gilt:

$\text{SUB}(w) :=$ Menge $V_{\text{Subgraphen}}$

$\text{EXT}(w) :=$ Menge $V_{\text{Extension}}$

(entsprechend der Beschriftung von w)

III. ESU-Baum

Notation

Seien die Knoten geordnet entsprechend der Reihenfolge der Funktionsaufrufe. Wird w_1 vor w_2 erzeugt, so schreiben wir $w_1 < w_2$.

Für eine Menge V von Knoten gilt:

Ein Knoten w_i heißt minimal, wenn er sich vor allen anderen Knoten w_j aus V befindet

III. ESU-Baum

Lemma 1:

1. Sei w_1 ein Knoten ungleich der Wurzel. Dann gilt für jeden Knoten $u \in \text{EXT}(w_1)$:
 w_1 hat ein Kind w_2 mit $u \in \text{SUB}(w_2)$
2. Für jeden Knoten w im ESU-Baum ungleich der Wurzel und jeden Knoten $u \in \text{EXT}(w)$ gibt es ein $v < u$, so dass gilt:
 v ist der kleinste Knoten in $\text{SUB}(w)$

III. ESU-Baum

Lemma 1 (*Fortsetzung*):

3. Seien w_1 und w_2 zwei Knoten mit dem gleichen Vater und $w_1 < w_2$. Dann enthält $\text{SUB}(w_1)$ genau einen Knoten u_1 , der nicht in $\text{SUB}(w_2)$ enthalten ist und umgekehrt. Für jeden Knoten w' , dessen Pfad zur Wurzel w_2 enthält, gilt: $u_1 \notin \text{SUB}(w')$

III. ESU-Baum

Beweis Lemma 1.1:

- Folgt direkt aus den Zeilen [E2] bis [E5]
- Für jeden Knoten $u \in V_{\text{Extesion}}$ wird ein Aufruf der Prozedur `ExtendSubgraph` ($V_{\text{Subgraph}} \cup \{u\}$, $V'_{\text{Extension}}, v$) erzeugt.
- Somit wird für jeden Knoten $u \in V_{\text{Extesion}}$ ein Kindsknoten erzeugt

III. ESU-Baum

Beweis Lemma 1.2:

- Folgt direkt aus den Zeilen [02] und [E4]
- Eine Bedingung, damit ein Knoten zu V_{Subgraph} hinzugefügt wird, ist dass er größer ist als v .
- Daraus folgt, dass der kleinste Knoten v ist, welcher als erstes eingefügt wurde, da dann nur noch größere Knoten eingefügt werden.

III. ESU-Baum

Beweis Lemma 1.3:

1. Fall: Der gemeinsame Vater ist die Wurzel im ESU-Baum

=> Zeile [03] wird genau einmal für jeden Knoten von Graph G ausgeführt.

Aus Bedingung 1.2 folgt, dass für jeden Nachfolger w' von w_2 gilt

$$\text{SUB}(w') \cap \text{SUB}(w_1) = \emptyset$$

III. ESU-Baum

Beweis Lemma 1.3 (*Fortsetzung*):

2. Fall: Der gemeinsame Vater ist ungleich der Wurzel im ESU-Baum

=> u_1 existiert, da er genau einmal zu V_{Subgraph} hinzugefügt wird und gleichzeitig aus V_{Extesion} entfernt wird [E3]

III. ESU-Baum

Beweis Lemma 1.3 (*Fortsetzung*):

Behauptung:

Nach dem Aufruf von $\text{ExtendSubgraph}(V_{\text{Subgraph}} \cup \{u_1\}, V_{\text{extension}}, v)$ [E5], wird kein Subgraph ausgegeben, der u_1 enthält, bis die Zeile E6 ausgeführt wird.

Beweis:

- u_1 ist Nachbar von einem Knoten aus V_{Subgraph} , solange er in $V_{\text{Extension}}$ ist.
 - Außerdem existiert kein Knoten u' aus V mit $u_1 \in N_{\text{excl}}(u', V_{\text{Subgraph}})$
- $\Rightarrow u_1$ wird nicht mehr zu $V_{\text{Extension}}$ hinzugefügt bis zu [E6]

III. ESU

Theorem ESU:

Gegeben sei ein Graph G und ein Integer $k \geq 2$.
ESU erzeugt alle Subgraphen der Größe k genau einmal.

III. ESU

Beweis Theorem ESU:

Mindestens einmal:

Annahme: Es existiert ein Subgraph G' der Größe k mit $\{v_1, \dots, v_k\}$, der nicht erzeugt wird.

O.B.d.A: Sei v_1 der kleinste Knoten in G'

=> Die Wurzel hat genau ein Kind w_1 mit

$\text{SUB}(w_1) = \{v_1\}$ [01-03]

III. ESU

Beweis Theorem ESU (Fortsetzung):

- Alle Nachbarn von v_1 sind in $\text{EXT}(v_1)$ [2].
- Aus 1.1 $\Rightarrow w_1$ hat ein Kind w_2 mit $\text{SUB}(w_2) = \{v_1, v'\}$ für jeden Nachbar v' von $v_1 \in G'$
- Sei w'_2 das minimale Kind
- O.B.d.A sei v_2 der gewählte Nachbar, so dass gilt $\text{SUB}(w'_2) = \{v_1, v_2\}$
- **Behauptung:** Alle Nachbarn von v_1 und v_2 sind in $\text{EXT}(w'_2)$ enthalten.

III. ESU

Beweis Theorem ESU (Fortsetzung):

Warum könnte diese Behauptung falsch sein?

1. Der Knoten könnte kleiner sein als v_1
(ABER: v_1 wurde als kleinster Knoten gewählt)
2. Kein Nachbar von v_1 und nicht in $\text{Nexcl}(v_2, \{v_1\})$.
(ABER: in G' ist er ein Nachbar von v_1 , v_2 oder beiden)
3. Könnte schon aus $\text{EXT}(w'_2)$ entfernt sein
(ABER: w'_2 wurde minimal gewählt)

III. ESU

Beweis Theorem ESU (Fortsetzung):

Wendet man diese Regeln induktiv für die Knoten v_3, \dots, v_k an, so kommt man zu einem Blatt w_k mit

$$\text{SUB}(w_k) = V(G')$$

III. ESU

Beweis Theorem ESU (Fortsetzung): Höchstens einmal:

Annahme: ESU erzeugt in den Blättern w_1 und w_2 den selben Subgraphen zweimal.

$\Rightarrow \text{SUB}(w_1) = \text{SUB}(w_2)$

- Die zugehörigen Pfade p_1 und p_2 müssen sich unterscheiden.
- Für den tiefsten gemeinsamen Knoten gilt: $\text{SUB}(w_1)$ und $\text{SUB}(w_2)$ unterscheiden sich mindestens in einem Knoten (Lemma 1.3)

III. ESU

Eigenschaften:

- Anzahl der Subgraphen der Größe k kann gut geschätzt werden.
- Laufzeit kann gut abgeschätzt werden
- Fehler können statistisch geschätzt werden (bei RAND-ESU)
- erzeugt unbiased, uniforme, zufällige Subgraphen (RAND-ESA)

III. RAND-ESU

Idee:

Anstatt den ganzen Baum zu erzeugen, können zufällig Teilbäume ausgelassen werden, so dass jedes Blatt mit der gleichen Wahrscheinlichkeit erreicht wird.

Sei $0 < p_d \leq 1$ eine Wahrscheinlichkeit für jede Tiefe $1 \leq d \leq k$ im ESU – Baum

p_d gibt an, mit welcher Wahrscheinlichkeit ein Knoten der Tiefe d erkundet wird (Teilbaum)

III. RAND-ESU

Eingabe : $G=(V, E)$ und $1 \leq k \leq |V|$, $p_d \forall 1 \leq d \leq k$

Ausgabe : Menge R von Subgraphen der Größe k

01 *for alle Knoten* $v \in V$ *do*

02 $V_{Extension} \leftarrow \{u \in N(\{v\}) : u > v\}$; $p_d := 1$

03 *Mit* p_d *ExtendSubgraph*($\{v\}$, $V_{Extension}$, v)

04 *return*

E1 *if* $|V_{Subgraph}| = k$ *then Ausgabe* $G[V_{Subgraph}]$ *und return*

E2 *while* $V_{Extension} \neq \emptyset$ *do*

E3 *Entferne einen zufälligen Knoten* w *aus* $V_{Extension}$

E4 $V'_{Extension} \leftarrow V_{Extension} \cup \{u \in N_{excl}(w, V_{Subgraph}) : u > v\}$

E5a $d := |V_{Subgraph}| + 1$

E5 *Mit* p_d *ExtendSubgraph*($V_{Subgraph} \cup \{w\}$, $V'_{Extension}$, v)

E6 *return*

III. RAND-ESU

Lemma 3:

Jedes Blatt im ESU-Baum wird mit der Wahrscheinlichkeit $\prod_d p_d$ besucht.

Beweis:

$$\begin{aligned} Pr[w_k \text{ besucht}] &= Pr[w_k \text{ besucht} | w_{k-1} \text{ besucht}] \\ &\quad * Pr[w_{k-1} \text{ besucht}] \\ &= p_k * Pr[w_{k-1} \text{ besucht}] \\ &= p_k * p_{k-1} * Pr[w_{k-2} \text{ besucht}] \\ &= \dots = \prod_{1 \leq d \leq k} p_d \end{aligned}$$

III. RAND-ESU

Satz 4:

Gegeben sind ein Graph G , ein Integer k und $0 < p_d \leq 1$ für alle $1 \leq d \leq k$. Sei R eine Menge von Subgraphen der Größe k , zufällig erzeugt von RAND-ESU. Dann ist

$$\hat{C}_k^i := \frac{|\{G' \in R : G' \in S_k^i(G)\}|}{|R|}$$

ein unbiased Schätzer für $C_k^i(G)$.

III. RAND-ESU

Beweis Satz 4:

Folgt aus Lemma 3:

Hat der Eingabegraph genau N Subgraphen und N' davon gehören zur Subgraph-Klasse S_k^i . Dann ist der Anteil der Blätter im ESU-Baum für $S_k^i = N'/N$. Da jedes Blatt mit gleicher Wahrscheinlichkeit besucht wird, gilt für den erwarteten Anteil von Subgraphen in R , die zu S_k^i gehören: $N'/N = C_k^i(G)$.

III. RAND-ESU

Wie sollten die Wahrscheinlichkeit p_d gewählt werden, wenn wir einen bestimmten Anteil $0 < q < 1$ erhalten wollen?

Es muss gelten: $\prod_{1 \leq d \leq k} p_d = q$

III. RAND-ESU

Beobachtungen:

- alle p_d gleich groß $\Rightarrow p_d = q^{\frac{1}{k}}$
- Je näher der Knoten an der Wurzel liegt, desto größer ist der Einfluss auf die Anzahl der besuchten Blätter, wenn ein Teilbaum untersucht wird, oder nicht.
- Ist p_d klein für kleine d , so werden wahrscheinlich einige lokale Nachbarn nicht besucht, während andere sehr ausgiebig besucht werden.

IV. Verfahren

1. Finde alle vorkommenden Subgraphen und deren Häufigkeit
2. Bilde Subgraph-Klassen, so dass alle topologisch gleichen Subgraphen in der selben Subgraph-Klasse liegen
3. **Bestimme welche Subgraph-Klassen besonders häufig auftreten im Vergleich zu einem bestimmten Zufallsgraph Modell**

IV. Signifikanz von Motiven

Explicit:

- Vergleiche die Konzentration C_k^i mit der durchschnittlichen Konzentration $\langle C_k^i(G) \rangle$ im Zufallsgraph
- Um $\langle C_k^i(G) \rangle$ zu schätzen, werden in der Regel ca. 1000 Zufallsgraphen mit gleicher Grad Sequenz erzeugt
- Zufallsgraphen mit gleicher Grad Sequenz werden erzeugt, in dem man den Graphen G nimmt und zufällig Kanten vertauscht

IV. Signifikanz von Motiven

Direct:

- Zufallsexperiment: Wähle eine Menge von k Knoten und bestimme dann die Menge der Graphen mit gleicher Grad Sequenz, in denen diese Knoten ein Subgraph von $S_k^i(G)$ bilden.

IV. Signifikanz von Motiven

Direct:

$$\langle C_k^i(G) \rangle$$

\approx

$$\langle \hat{C}_k^i(G) \rangle$$

$=$

$$\sum_{\{v_1, \dots, v_k\} \subseteq V} \left| \{ G' \in SEQ(G) : G'[v_1, \dots, v_k] \in S_k^i \} \right|$$

$$\sum_{\{v_1, \dots, v_k\} \subseteq V} \left| \{ G' \in SEQ(G) : G'[v_1, \dots, v_k] \text{ is connected} \} \right|$$

IV. Signifikanz von Motiven

Theorem 5 (ungerichtete Graphen):

$$G(M, r) \sim \frac{\sqrt{2} \left(\frac{f}{e}\right)^{\left(\frac{f}{2}\right)}}{\exp(a^2 + a + b) * \prod_i (r_i!)}$$

$$\text{mit } f \rightarrow \infty, a = \sum_i \frac{r_i^2 - r_i}{2f}, b = \sum_{m_{ij}=0, i < j} \frac{r_i * r_j}{f}, f = \sum_i r_i$$

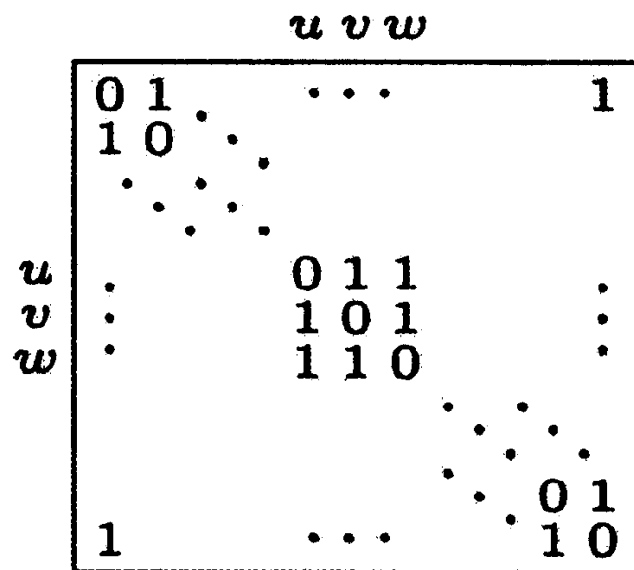
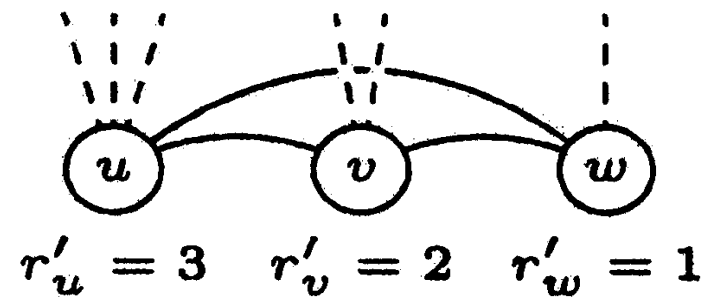
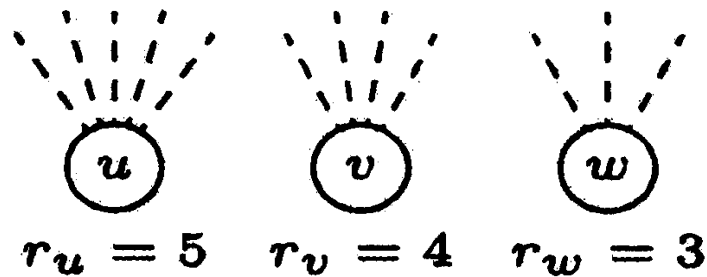
IV. Signifikanz von Motiven

Idee:

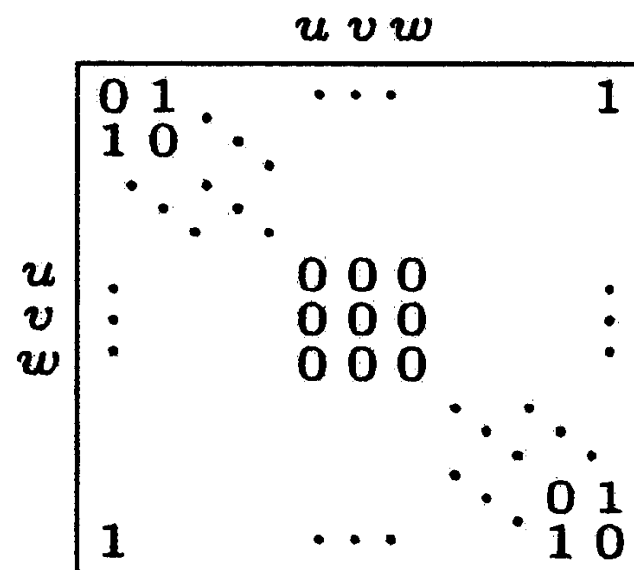
- Zähle die Anzahl Graphen mit gleicher Grad Sequenz, die alle ein bestimmten Subgraph enthalten.
- Um einen Kante zwischen 2 bestimmten Knoten zu verbieten, setze die entsprechenden Einträge in der Bitmask Matrix auf Null.
- Um eine Kante zu erzwingen reduziere zusätzlich den Grad der beiden Knoten um eins.

IV. Signifikanz von Motiven

Bitmask-Matrix:



M



M'

IV. Signifikanz von Motiven

Direct:

$$\langle C_k^i(G) \rangle$$

\approx

$$\langle \hat{C}_k^i(G) \rangle$$

$=$

$$\sum_{\{v_1, \dots, v_k\} \subseteq V} \left| \{ G' \in SEQ(G) : G'[v_1, \dots, v_k] \in S_k^i \} \right|$$

$$\sum_{\{v_1, \dots, v_k\} \subseteq V} \left| \{ G' \in SEQ(G) : G'[v_1, \dots, v_k] \text{ is connected} \} \right|$$

IV. Signifikanz von Motiven

Schätzen des Zählers:

Sei eine Subgraph-Klasse S_k^i und eine Knotenmenge $\{v_1, \dots, v_k\}$ gegeben.

=> Es existieren $k!$ Möglichkeiten, auf welche die Knotenmenge $\{v_1, \dots, v_k\}$ einen Subgraph für S_k^i induzieren kann.

IV. Signifikanz von Motiven

Schätzen des Nenners:

Haben wir den Zähler für alle Subgraph-Klassen S_k^i bestimmt, so ist der Nenner die Summe über alle Zähler.

Wurde nur ein Teil der Subgraph-Klassen bestimmt, so fordern wir nur, dass die Knoten $\{v_1, \dots, v_k\}$ verbunden sind, ohne einen festen Subgraph festzulegen.

IV. Signifikanz von Motiven

Schätzen des Nenners (Fortsetzung):

Dies erreichen wir, in dem wir einen Spannbaum zwischen den Knoten $\{v_1, \dots, v_k\}$ erzwingen. Um nicht manche Subgraphen doppelt zu zählen, versehen wir die Kanten mit paarweise verschiedenen Gewichten:

Für alle $\{v_i, v_j\}$ mit $i < j : i+k*j$ ist Kantengewicht

IV. Signifikanz von Motiven

Schätzen des Nenners (Fortsetzung):

Theorem 6:

Jeder Graph mit paarweise verschiedenen Kantengewichten, hat einen eindeutigen minimalen Spannbaum.

IV. Signifikanz von Motiven

Schätzen des Nenners (Fortsetzung):

Somit müssen wir $k^{(k-2)}$ Bäume betrachten, um den Nenner zu schätzen.

Jeder von diesen Bäumen muss einen minimalen Spannbaum auf der Knotenmenge $\{v_1, \dots, v_k\}$ haben, sonst wird er nicht berücksichtigt.