Poznámka
Úvodní přednáška.

Definice 0.1

Průměrná nejkratší cesta:

$$L(G) = \frac{1}{n(n-1)} \sum_{u \neq v} d(u, b).$$

1 Centrality

Definice 1.1 (Ekcentricita)

Excentricita je pro každý vrchol největší vzdálenost do jiného vrcholu.

$$\varepsilon(v) = \max_{u \in V(G)} d_G(v, u).$$

Definice 1.2 (Closeness centralita)

Closeness centralita je převrácená hodnota průměrné vzdálenosti do vrcholů.

$$C_C(v) = \frac{n}{\sum_u d(u, v)}.$$

Definice 1.3 (Betweenness centralita)

Betweenness centralita je poměr počtu cest, které vedou skrz daný vrchol vůči všem cestám sečteno přes všechny dvojice vrcholů.

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma(st)},$$

kde $\sigma_{st}(v)$, resp. σ_{st} je počet nejkratších cest z vrcholu s do vrcholu t jdoucích přes v, resp. všech.

Definice 1.4 (Globální betweenness)

$$C_B(G) = \frac{1}{n} \sum_v C_B(v).$$

Věta 1.1

Mějme graf o n vrcholech. Potom

$$C_B(G) = (n-1)(L(G) - 1).$$

Důkaz
Vezmeme libovolné dva body a představíme si nejkratší cesty mezi nimi a každou „vlnu“ (tj. všechny vrcholy na nejkratších cestách ve vzdálenosti 1, 2, 3, …). Tyto dva body nám do každé vlny přispějí v součtu 1 (krát $1/n$). Tedy dva body do globální betweenness přispějí průměrně $L(G) - 1$ (krát $1/n$). Dvojic je $n \cdot (n-1)$, ale protože globální betweenness se dělí ještě n, tak výsledek je $(n-1)(L(G)-1)$. □

Důsledek
S klesající průměrnou nejkratší vzdáleností (tedy např. přidáním hrany) klesá i betweenness.

Lemma 1.2

Přidáním hrany mezi vrcholy u, v, $d = d_G(u,v) > 1$ snížíme globální betweenness nejméně o $\frac{2(d-1)}{n}$.

Důkaz
$L(G)$ se sníží alespoň o $\frac{2(d-1)}{n(n-1)}$, protože ze dvou vrcholů ve vzdálenosti d uděláme vrcholy ve vzdálenosti 1. Odečtením jedničky a vynásobením $(n-1)$ dostáváme, ze $C_B(G) = (n-1)(L(G)-1)$ se nám sníží alespoň o $\frac{2(d-1)}{n}$. □

Důsledek
Globální betweenness grafu je minimálně o $\frac{2r}{n}$ menší než globální betweenness jeho kostry, kde r je rozdíl mezi počtem hran grafu a jeho kostry, tj. $m - n + 1$.

Definice 1.5 (k-okolí vrcholu)

k-okolí vrcholu v je definováno jako

$$\Gamma_k(v) = \{u \in V : d(u,v) = k\}.$$

Věta 1.3

Buď G graf velikosti n, $w \in V$ jeho vrchol a G' graf získaný z G přidáním vrcholu u a hrany $\{u, w\}$. Potom

$$C_B(G) = \frac{n}{n+1} C_B(G) + \frac{2}{n+1} \sum_{k=1}^{\varepsilon w} k|\Gamma_k(w)|.$$

Důkaz

Původní vzdálenosti se vůbec nezmění, jen už nebudou průměrovaný přes $n \cdot (n-1)$, ale přes $(n+1) \cdot n$. Navíc přidáváme cesty (každou dvakrát, „jednou sem a jednou tam“) z nového vrcholu do všech dalších a ty musí vést přes w a tedy

$$L(G') = \frac{n-1}{n+1}L(G) + \frac{2}{n(n+1)}\sum_{k=1}^{\varepsilon(w)}(k+1)|\Gamma_k(w)| = \frac{n-1}{n+1}L(G) + \frac{2}{n(n+1)}\left(\sum_{k=1}^{\varepsilon(w)} k|\Gamma_k(w)| + n\right),$$

$$C_B(G') = n(L(G')-1) = \frac{n}{n+1}(n-1)(L(G)-1) + \frac{n(n-1)}{n+1} + \frac{2}{n+1}\sum_{k=1}^{\varepsilon(w)} k|\Gamma_k(w)| + \frac{n(1-n)}{n+1} = \frac{n}{n+1}(n-$$

$\square$

Tvrzení 1.4

Pokud do grafu G o n vrcholech přidáme vrchol w s 2 hranami do vrcholů u, v v původním grafu ve vzdálenosti $1 \leq d(u,v) \leq 2$ a získáme tím G', tak

$$\frac{1}{n+1}(nC_B(G) + n - 2) \leq C_B(G').$$

Důkaz

Pokud je vzdálenost 1, pak u, v, w tvoří trojúhelník, $C_B(w) = 0$ a všechny cesty z w vedou přes u nebo v. Tedy máme $n-2$ vrcholů, který spolu s w přispějí do C_B v a w v součtu hodnotou $n-2$. Do ostatních C_B přispějí něčím. Navíc musíme průměrovat přes $n+1$ vrcholů místo n, tedy dostáváme

$$\frac{1}{n+1}(nC_B(G) + n - 2 + \text{něco}) = C_B(G').$$

Pro vzdálenost $d(u,v) = 2$ se C_B z vrcholů na cestách mezi u, v rozdělí navíc do W, ale součet zůstane stejný. Jinak to vypadá jako v prvním případě. $\square$

Tvrzení 1.5

$$0 \leq C_B(v) \leq (n-1)(n-2).$$

Důkaz

Triviální. C_B z konkrétní uspořádané dvojice dostane přispěno maximálně 1. $\square$

Důkaz

$$C_B(v) = 0 \Leftrightarrow G[\Gamma_G(v)] \cong K_{\deg(v)}.$$

Kde $\Gamma_G(v)$ je množina sousedů bodu v a $K_{\deg(v)}$ je úplný graf o $\deg(v)$ vrcholech.

Důkaz
Triviální. □

□

Poznámka
Graf s vrcholem, který má libovolnou danou betweenness lze sestrojit ze 2 úplných grafů (spojených do 1 úplného) a 2 vrcholů (spojených se sebou a každý se svým úplným grafem)

Tvrzení 1.6

Mějme graf G s maximálním stupněm δ. Potom

$$C_{B\max}(G) = \max_{v \in V(G)} C_B(v) \leq \frac{\delta - 1}{2\delta}(n-1)^2.$$

Důkaz
Nechť w je vrchol s maximální C_B a stupněm $\deg(w) = d \leq \delta$. Spustíme prohledávání do hloubky z w a získáme tak nějakou kostru T. Pro každého souseda w_i vrcholu w máme podstrom T_i, $t_i = |T_i|$. Nejkratší cesty vedou mezi stromy přes w a uvnitř stromů mimo w. Tím, že přidáme hrany zpět jen snížíme počet cest vedoucích přes w, tedy

$$C_B(w) \leq \sum_{i,j=1 i \neq j}^{d} t_i t_j \leq \binom{\delta}{2} \left(\frac{n-1}{\delta}\right)^2 = \frac{\delta - 1}{2\delta}(n-1)^2,$$

jelikož suma nabývá maximum pro $d = \delta$ a $t_i = t_j$. □

Definice 1.6

$C_{B\max}(G)$ tedy značí maximální C_B ze všech vrcholů grafu. Definujeme ještě $C_{B\max}(\mathcal{H})$ jako maximum z maximálních betweenness ze všech grafů v třídě $\mathcal{H}$.

Tvrzení 1.7

Pro 2-connected graf G velikosti n je:

$$C_{B\max}(G) \leq \frac{(n-3)^3}{2}.$$

Rovnost nastává pro $F_{1,n-1}$.

Důkaz
TODO. $\square$

Poznámka
Automorfismy zachovávají centrality. Ale i další zobrazení mohou zachovávat centrality.

Definice 1.7 (Betweenness uniformní graf)

Betweenness uniformní graf je takový graf, který má C_B každého vrcholu stejnou.

Věta 1.8

Každý betweenness uniformní graf je 2-connected.

Důkaz
Sporem. TODO. $\square$

Lemma 1.9

Buď $G = (V, E)$ graf s průměrem 2. Potom $\forall x \in V$:

$$C_B(x) = \sum_{(u,v)\in\Gamma^2(x) uv\notin E(G)} 1/\sigma_{u,v}.$$

Důkaz
Triviálně z definice. $\square$

Lemma 1.10

Buď $G = (V, E)$ betweenness uniformní graf s průměrem 2. Potom $\forall x \in V : C_B(x) = n - 1 - 2m/n$.

Důkaz
Z předchozího lemmatu. TODO. $\square$

Věta 1.11

Buď $G = (V, E)$ betweenness uniformní graf s $\delta(G) = n - 1$ a $|V| = n$. Potom $G \cong K_n$.

Důkaz
Buď $x \in V(G)$, $\deg(x) = n - 1$ a $H = G - x$. TODO. $\square$

Věta 1.12

Buď $G = (V, E)$ betweenness uniformní graf, kde $\delta(G) = n - 2$. Potom $\operatorname{diam} G = 2$.

Věta 1.13

Buď $G = (V, E)$ betweenness uniformní graf. Potom $C_B(v) = 0$ nebo $C_B(v) \geq 1$.

Definice 1.8 (Degree centralita)

Degree centralita je normalizovaný stupeň daného vrcholu:

$$\frac{\deg(v)}{n-1}.$$

Definice 1.9 (Eigenvector centralita)

$$\mathbf{x}(i) = A\mathbf{x}(i-1) \Leftrightarrow \mathbf{x}(i) = A^i\mathbf{x}(0), i \to \infty.$$

Kde A je matice sousednosti.

Lemma 1.14 (Spektrální vlastnosti matice sousednosti)

Buď A matice sousednosti neorientovaného grafu a $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ její vlastní čísla. Potom $\sum \lambda_i = 0$, $\lambda_1 \geq 0$, $\det(A) = \prod \lambda_i$.

Lemma 1.15 (Pozitivita vlastního páru)

Buď A matice sousednosti a $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ její vlastní čísla. Potom $\lambda_1 \geq |\lambda_i|$ a příslušný vlastní vektor má všechny složky nezáporné.

Definice 1.10 (Katzova centralita)

Jako eigenvector centralita, jen přidáme konstantu, aby nebyl problém s orientovanými grafy:

$$x_i = \alpha \sum A_{ij}x_j + \beta \Leftrightarrow \mathbf{x} = \alpha AA\mathbf{x} + \beta\mathbf{1}.$$

Pro $\beta = 1$ vychází:

$$\mathbf{x} = (I - \alpha A)^{-1}\mathbf{1}$$

Poznámka
Katzova centralita má problém, že preferuje sousedy vrcholů s velkým stupněm a velkou centralitou. Proto se používá:

Definice 1.11 (Pagerank)

Jako Katzova centralita, ale dělí příspěvky od vrcholů jejich výstupními stupni (respektive 1, když má stupeň 0). Dostáváme tedy

$$x_i = \alpha \sum A_{ij} \frac{x_j}{\max\left\{\deg^+(j), 1\right\}} + \beta$$

a při volbě $\beta = 1$ máme:

$$\mathbf{x} = (I - \alpha A D^{-1})\mathbf{1}.$$

Poznámka
Pro neorientované grafy se α volí jako 1, pro orientované musí být jiná. Např. Google používá 0.85.

Definice 1.12 (Wienerův index)

$$W(G) = \sum_{\{u,v\} \subseteq V(G)} d(u,v).$$

Používá se například k určování bodu varu molekul.

2 Náhodné grafy

Definice 2.1 (Náhodný binomický graf $G_{n,p}$)

2 parametry: počet vrcholů n a pravděpodobnost hrany p.

Definice 2.2 (Náhodný uniformní graf $G_{n,m}$)

2 parametry: počet vrcholů n a počet hran m.

Definice 2.3 (Průměrný počet hran)

$$< m >= \sum_{i=1}^{\binom{n}{2}} 1 \cdot p = p\binom{n}{2}.$$

Definice 2.4 (Průměrný stupeň)

$$< k >= \left\langle \frac{2m}{n} \right\rangle = \frac{2 < m >}{n} = p(n-1).$$

Definice 2.5 (Řídký graf)

Graf nazýváme řídký, pokud $< k ><< n$.

Definice 2.6 (Pravděpodobnost stupně k pro daný vrchol)

$$p_k = \binom{n-1}{k} p^k (1-p)^{n-1-k}.$$

Pro řídké grafy

$$p_k = \frac{< k >^k}{k!} e^{-<k>}.$$

Tedy Poissonovo rozdělení pro $\lambda =< k >$.

2.1 Obrovská komponenta

Definice 2.7 (Obrovská komponenta (GC))

Komponentu C grafu $G = (V, E)$ nazýváme obrovská, pokud $|C| \sim |V|$.

Lemma 2.1

Pravděpodobnost, že $v \in V(G)$ není k GC připojen vrcholem $w \in V(G)$ je $1 - p + pu$, kde u je počet vrcholů mimo obrovskou komponentu (normovaný k celému počtu, tedy pravděpodobnost, že náhodný vrchol nebude součástí GC).

Důkaz

2 možnosti, jak nebýt připojen přes vrchol 2: Pravděpodobnost, že mezi nimi nevede hrana je $1 - p$ a pravděpodobnost, že vede a w není součástí GC je $p \cdot u$. Tedy $1 - p + pu$. $\square$

Věta 2.2

Buď $G = (V, E)$ graf s průměrným stupněm $< k >$. Potom G má obrovskou komponentu $\Leftrightarrow < k >> 1$.

Důkaz
Pravděpodobnost, že v není připojeno přes w do GC je $1 - p + pu$, tedy že není připojen žádným je $(1 - p + pu)^{n-1} = u$. Substituce $< k >= p(n-1)$ nám dává:

$$u = \left(1 - \frac{< k >}{n-1}(1-u)\right)^{n-1}, \qquad \ln u = (n-1)\ln\left(1 - \frac{< k >}{n-1}(1-u)\right) \approx -(n-1)\frac{< k >}{n-1}(1-u) = - <$$

Tedy máme aproximaci a pro velké grafy dokonce rovnost $\ln u =< k > (1-u)$, tj. při $S = 1 - u$ máme $S = 1 - e^{<k>S}$. Graficky odhadneme, že pro $< k >\leq 1$ je $S \approx 0$. Dále rozebereme, jak rostou l-okolí nějakého bodu s rostoucím l a dostaneme se k tomu, že pro $< k >> 1$ rostou exponenciálně, tedy tvoří GC. □

Věta 2.3

Počet obrovských komponent v řídkých grafech je 1, tj. pravděpodobnost 2 a více GC konverguje k 0.

Důkaz
TODO. □

Definice 2.8 (Malá komponenta)

Komponenta C je malá, pokud $|C| << n$.

Pozorování
Ať C je malá komponenta $G_{n,p}$. Potom s velkou pravděpodobností je izomorfní stromu.

Důkaz
TODO. □

Pozorování
Náhodný graf $G_{n,p}$ s m hranami je se stejnou pravděpodobností jeden ze všech grafů o n vrcholech a m hranách.

Důkaz
Triviální. □

2.2 Malý svět (small-world)

Definice 2.9 (Koeficient klusterizace)

$$C = \frac{\#\text{tojúhelníků}}{\#\text{spojených trojic}} = \frac{\binom{n}{3}p^3}{\binom{n}{3}p^2} = p = \frac{< k >}{n-1}.$$

Poznámka
V náhodných grafech můžeme počet vrcholů ve vzdálenosti d vyjádřit jako

$$N(d) \approx 1 + < k > + < k >^2 + \ldots + < k >^d = \frac{< k >^{d+1} - 1}{< k > - 1} \approx < k >^d .$$

Definice 2.10 (Malý svět)

Graf má vlastnost malého světa, pokud $L \gtrsim L_{rand}$ a $C >> C_{rand}$, tedy průměrná délka cest je alespoň taková jako u náhodného grafu a koeficient klusterizace je řádově větší než u náhodného grafu.

Zavádíme také koeficient small-world a podmínky výše můžeme znázornit jako:

$$\sigma = \frac{\frac{L}{L_{rand}}}{\frac{C}{C_{rand}}} >> 1.$$

3 Vlastnosti grafu

Definice 3.1

Grafová vlastnost $\mathcal{P}$ je podmnožina všech grafů (s označenými vrcholy) na n vrcholech, tj. $\mathcal{P} \subseteq 2^{\binom{V}{2}}$.

Například
Souvislé grafy, rovinné grafy, grafy neobsahující K_3, ...

Definice 3.2

O vlastnosti grafu říkáme, že je monotónní, pokud ji graf přidáním hrany neztratí, tj. $\forall G \in \mathcal{P} \forall e \in \binom{V}{2} : G + e \in \mathcal{P}$.

Lemma 3.1

Buď $\mathcal{P}$ monotónní vlastnost grafu a $0 \leq p_1 < p_2 \leq 1$. Potom

$$\mathbf{P}[G_{n,p_1} \in \mathcal{P}] \leq \mathbf{P}[G_2 \in \mathcal{P}].$$

Důkaz
Vytvoříme graf G tak, aby $G_{n,p_2} = G_{n,p_1} \cup G$. Lehce si rozmyslíme, že $G = G_{n,p}$, kde $p = 1 - \frac{1-p_2}{1-p_1}$. □

Lemma 3.2

Buď $\mathcal{P}$ grafová vlastnost a $p = \frac{m}{\binom{n}{2}}$, kde $m = m(n) \to \infty$ a $\binom{n}{2} - m \to \infty$. Potom pro velká n je

$$\mathbf{P}[G_{n,m} \in \mathcal{P}] \leq \sqrt{2\pi m}\mathbf{P}[G_{n,p} \in \mathcal{P}].$$

Důkaz
TODO. (Přes Stirlingovu formuli dokážeme $\mathbf{P}[E_{n,p} = m] \geq \frac{1}{\sqrt{2\pi m}}$.) □

Lemma 3.3

Buď $\mathcal{P}$ monotónní grafová vlastnost a $p = \frac{m}{N}$, kde $N = \binom{n}{2}$. Potom pro velká n a p máme

$$\mathbf{P}[G_{n,m} \in \mathcal{P}] \leq 3\mathbf{P}[G_{n,p} \in \mathcal{P}].$$

Důkaz
Rozepíšeme si pravděpodobnost po počtech hran:

$$\mathbf{P}[G_{n,p} \in \mathcal{P}] \geq \mathbf{P}[G_{n,m} \in \mathcal{P}] \sum_{m'=0}^{N} \mathbf{P}[E_{n,p} = m'].$$

Potom s tím uděláme magii a dostaneme dolní odhad na sumu rovný $\frac{1}{3}$. □

Definice 3.3 (Threshold)

Funkce p^* je threshold pro monotónní vlastnost $\mathcal{P}$ v náhodném grafu $G_{n,p}$ pokud

$$\lim_{n\to\infty} \mathbf{P}[G_{n,p} \in \mathcal{P}] = 0, \text{ pokud } p/p^* \to 0, \text{ a } 1, \text{ pokud } p/p^* \to \infty.$$

Lemma 3.4

Ať $\mathcal{P}$ je grafová vlastnost říkající, že graf vlastní alespoň 1 hranu. Potom

$$\lim_{n\to\infty} \mathbf{P}[G_{n,p} \in \mathcal{P}] = 0, \textit{ pokud } p >> n^{-2}, \textit{ a } 1, \textit{ pokud } p << n^{-2}.$$

Důkaz
Triviální, s využitím $\mathbf{P}[X > 0] \leq \mathbb{E}X$, kde X je počet hran v grafu, máme $\mathbf{P}[X > 0] \leq \frac{n^2}{2}p$. A využitím $\mathbf{P}[X = 0] \leq \frac{VarX}{(\mathbb{E}x)^2}$ máme $\mathbf{P}[X > 0] \geq 1 - \frac{1-p}{\binom{n}{2}p}$. ☐

Věta 3.5

Pokud $m/n \to \infty$, potom asymptoticky téměř jistě $G_{n,m}$ obsahuje alespoň jeden trojúhelník.

Důkaz
Protože vlastnost „obsahuje alespoň 1 trojúhelník" je monotónní, tak to dokážeme pro $G_{n,p}$ a $np \to \infty$. Buď $\mathcal{P}_Z$ vlastnost, že počet trojúhelníků je 0. Potom dokážeme $\mathbf{P}[G_{n,p} \in \mathcal{P}_Z] \to 0$ a $\mathbf{P}[G_{n,m} \in \mathcal{P}_Z] \leq 3\mathbf{P}[G_{n,p} \in \mathcal{P}_Z]$???. TODO! (Asi 6 slidů.) ☐

Definice 3.4 (Evoluce grafů)

Definujeme dynamický proces $G_0 = ([n], \emptyset), G_1, G_2, \ldots, G_N = K_n$, kde $G_m \to G_{m+1}$ je provedeno pomocí přidání náhodné hrany.

Pozorování
G_m a $G_{n,m}$ mají stejné pravděpodobnostní rozdělení.

Věta 3.6

Pokud $m << n$, potom G_m je asymptoticky jistě les.

Důkaz
TODO! ☐

3.1 Modelování náhodných grafů

Poznámka
Problém náhodného rozdělení grafů je, že rozdělení stupňů je dáno Poissonovým rozdělením, kdežto v reálných grafech je dáno často scale-free rozdělením (malé stupně jsou nejčastější, pravděpodobnost stupně k je $p_k = Ck^{-\alpha}$, kde $2 \leq \alpha \leq 3$).

Definice 3.5 (Konfigurační model)

Mějme definovanou posloupnost stupňů $k_1, k_2, \ldots, k_n$. Vytvoříme „pahýly", tedy vrcholy s k_i „polohranami". Potom vezmeme náhodné dvě „polohrany" a spojíme.

Pozorování
Vzniknou nám dvojité hrany i smyčky. Pravděpodobnost dvojité hrany mezi vrcholy je $\frac{k_i k_j (k_i-1)(k_j-1)}{(2m)^2}$, tedy počet dvojitých hran bude přibližně

$$P_d = \frac{1}{2}\frac{1}{(2m)^2}\sum_{i,j} k_i k_j (k_i - 1)(k_j - 1) = \frac{1}{2}\left(\frac{< k^2 > - < k >}{< k >}\right)^2.$$

Počet smyček bude podobně přibližně

$$P_L = \sum_i \frac{k_i(k_i - 1)}{4m} = \frac{< k^2 > - < k >}{2 < k >}.$$

Definice 3.6 (Chung-Lu model)
Přiřadíme váhy (w_i = očekávaný stupeň) vrcholům. Následně pro $m := \sum_i w_i$ spojíme vrcholy i, j s pravděpodobností $p_{ij} = \frac{w_i w_j}{2m}$.

Pozorování
Průměrný počet hran je

$$\sum_{i<j} p_{ij} + \sum_i p_{ii} = \sum_{i<j} \frac{w_i w_j}{2m} + \sum_i \frac{w_i^2}{2m} = \sum_{i,j} \frac{w_i w_j}{4m} = \frac{4m^2}{4m} = m.$$

Průměrný stupeň vrcholu v_i je

$$< k_i >= 2p_{ii} + \sum_{i\neq j} p_{ij} = \sum_j \frac{w_i w_j}{2m} = w_i.$$

Definice 3.7 (Barabási-Albert model)
Začínáme s m_0 vrcholy propojenými libovolně (ale každý se stupněm > 0). Přidáme nový vrchol a propojíme ho s $m \leq m_0$ existujícími vrcholy tak, že

$$\mathbf{P}[u \text{ je spojený s existujícím vrcholem stupně } k_i] = \frac{k_i}{\sum_j k_j}.$$

Pozorování
Problémem je, že graf na m_0 vrcholech není daný, mohou vznikat multihrany, ale distribuce stupňů je $p_k = Ck^{-\gamma}$, kde $\gamma = 3$.

4 Komunity

Definice 4.1 (Komunita)

Podgraf C je silná komunita, jestliže pro každé $v \in C$ je počet hran vedoucích mimo C menší než počet hran vedoucích do C.

Podgraf C je slabá komunita, pokud počet hran vedoucích z C pryč je menší než počet hran uvnitř C.

Občas se uvažuje i vážená příslušnost ke komunitám...

Definice 4.2

Intra-cluster hustota je $\varrho_{int} := |E(C)|/\binom{n_c}{2}$.

Inter-cluster hustota je $\varrho_{ext} := |E(C, G \setminus C)|/(n_c(n - n_c))$.

Hustota grafu je $\varrho := |E(G)|/\binom{n}{2}$.

Důsledek
Komunita má $\varrho_{int} >> \varrho >> \varrho_{ext}$, takže při hledání komunit chceme maximalizovat $\varrho_{int} - \varrho_{ext}$.

Definice 4.3

k-klika je maximální podgraf, kde vzdálenost (v grafu G, ne přímo v klice) každých vrcholů je k. k-klub je k-klika s průměrem (už přímo v klubu, zajišťuje souvislost) nejvýše k.

k-plex je maximální podgraf, kde každý vrchol je spojený se všemi ostatními krom $k-1$ z nich.

Věta 4.1

k-PLEX je NP-úplný problém pro libovolné k.

Důkaz
Převedením CLIQUE na k-PLEX. Viz prezentace. $\square$

Poznámka
Více rozdělování do komunit se objevuje ve strojovém učení. Viz SUVPV.

Definice 4.4 (Girvan-Newman algoritmus)

Cílem je definovat hranovou centralitu tak, aby byla velká pro různé komunity a malá pro stejné. Na optimum se nabízí tzv. edge-betweenness centralita:

$$C_{EB}(e) = \sum_{u,v \notin e} \frac{\sigma_{uv}(e)}{\sigma_{uv}}.$$

Algoritmus tedy spočítá centralitu pro každou hranu, odstraní hranu s nejvyšší centralitou, přepočítá centrality a postupně takto odstraní všechny hrany, čímž dostaneme chtěnou hierarchii. (Pak zbývá jen otázka, kolik chceme komunit.)

Definice 4.5 (Modularita)

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j),$$

kde P_{ij} je tzv. null model, $\delta(C_i, C_j)$ je indikátor stejného klusteru.

Poznámka (Volba null modelu)
Prostě stejný počet hran: $P_{ij} = p = \frac{2m}{n(n-1)}$.

Model zachovávající stupeň: $P_{ij} = 2mp_i p_j = \frac{k_i k_j}{2m}$.

Poznámka (Na základě modelu zachovávající stupeň)

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j).$$

První část je:

$$\frac{1}{2m} \sum_{ij=1}^{n} A_{ij} \delta(C_i, C_j) = \frac{1}{2m} \sum_{c=1}^{n_c} \sum_{i,j \in C_c} A_{ij} = \sum_{c=1}^{n_c} \frac{m_c}{m}.$$

Druhá pak:

$$\frac{1}{2m} \sum_{ij=1}^{n} \frac{k_i k_j}{2m} \delta(C_i, C_j) = \sum_{c=1}^{n_c} \frac{d_c^2}{4m^2}.$$

Tedy modularita je:

$$Q = \sum_{c=1}^{n_c} \left(\frac{m_c}{m} - \left(\frac{d_c}{2m}^2 \right) \right).$$

Tudíž maximalizujeme $\sum_{c=1}^{n_c} [m_c - \mathbb{E}[m_c]]$.

Definice 4.6 (Kernighan-Lin algoritmus)

Pouze přibližné dělení: Začne s počátečním dělením (např. náhodným) X, Y a pro všechny dvojice vrcholů z opačných částí spočítá, jak by se zlepšilo $D_x = k_x^{ext} - k_x^{int}$ při jejich prohození. Nejlepší dvojici pak prohodí. Opakuje (ale neprohazuje již prohozené vrcholy), dokud se snižuje výsledek.

Věta 4.2

Souvislý graf má maximální vlastní číslo matice modularity rovné 0 právě tehdy, pokud je to úplný multipartitní graf.

Důkaz
TODO. □

Tvrzení 4.3

TODO.

Věta 4.4 (Petrovi c)

Matice souvislosti souvislého grafu má právě jedno kladné vlastní číslo tehdy, a jen tehdy, pokud je to úplný multipartitní graf.

Lemma 4.5

TODO.

Definice 4.7 (Vzdálenost)

Vzdálenost je funkce $d : S \times S \to \mathbb{R}$ splňující

$$\forall i, j \in S, i \neq j : d(i, j) \geq 0,$$

$$\forall i, j \in S : i = j \Leftrightarrow d(i, j) = 0.$$

Definice 4.8 (Klustrizace)

Klusterizace je funkce f z množiny všech vzdáleností na S do množiny funkcí z S na jeho rozdělení, tj. do podmnožiny $\mathcal{P}$.

Množiny dělení jsou nazvány klustery.

Definice 4.9 (Co splňuje dobré dělení)

Je invariatní vůči škálování: Pro každou vzdálenost a každé $\alpha > 0$ máme $f(d) = f(\alpha d)$.

Bohatost: Rang(f) je množina všech dělení S.

Konzistentnost: Když zmenšíme vzdálenosti uvnitř klusterů a zvětšíme mezi nimi, dělení zůstane stejné.

Definice 4.10 (Single-linkage dělení)

Začneme s klusterem za každý bod a vždy spojíme klustery, které jsou nejblíže, dokud není splněna ukončovací podmínka.

Pozorování
Pokud zastavíme na k klusterech, tak dělení není bohaté.

Pokud zastavíme na vzdálenosti r, tak dělení není invariantní vůči škálování.

Pokud zastavíme na α násobku maximální délky hrany, tak dělení není konzistentní.

Věta 4.6 (Kleinberg 2002)

Pro každé $n \geq 2$ neexistuje klusterizace f, která by splňovala všechny 3 podmínky.

Věta 4.7 (Kleinberg 2002)

Pokud klusterizace splňuje invariantnost vůči škálování a je konzistentní, pak $\mathrm{Rang}(f)$ *je antiřetězec.*

Důkaz
TODO! □

Věta 4.8 (Kleinberg 2002)

Pro každý antiřetězec dělení existuje klusterizace, která splňuje invarianci vůči škálování a konzistenci, která má obor hodnot přesně tento řetězec.

5 Šíření v sítích

Definice 5.1 (Stavy jedince)

S (susceptible) = zdravý, I (Infected) = nakažený, R (recovered) = uzdravený.

Definice 5.2 (Klasický SI model)

Předpokládá, že se každá dvojice potkává se stejnou pravděpodobností.

Má 2 parametry: β = pravděpodobnost I -> S (při setkání) a $< k >$ = průměrný počet kontaktů 1 osoby.

Definice 5.3 (SIS model)

SIS model navíc předpokládá i možnost znovunakažení.

Definice 5.4 (SIRS model)

SIRS model navíc přidává vyléčené, kteří se znovu nakazí s menší pravděpodobností.