

Využití grafů při analýze sociálních sítí

Michail Tsikerdekis

Grafy a sítě

podzim 2011

O čem bude dnešní přednáška

- Motivace
- Teorie grafů a sociální sítě
- Analýza sociálních sítí
- Ukázková studie
- Shrnutí

K čemu je analýza sociálních sítí?

- Člověk žije ve společenství, vytváří si různé vazby a vztahy s ostatními lidmi, které se zpravidla mění v čase
- Obvykle jsou sociální sítě vícevrstvé a lze se na ně dívat z různého úhlu pohledu a v různých kontextech
- Studium sociálních sítí využívá poznatků několika vědních oborů: psychologie, antropologie, sociologie a diskrétní matematiky (teorie grafů)
- **Šest stupňů odloučení** – teorie, která předpokládá, že každý člověk je spojený s libovolným jiným člověkem na Zemi prostřednictvím řetězce šesti sobě navzájem známých lidí

Historie

- 18. stol Leonhard Euler – položil základy teorie grafů
- 30. léta 20. stol. Jacob Moreno (sociometrie) – kvantitativní metodologie pro zkoumání sociálních vazeb
- 40. léta 20. stol. Formální definice *klik* v teorii grafů
- 50. léta 20. stol. Elizabeth Bott, John Barnes (antropologie) – Manchesterská škola, zasloužili se o rozvoj výzkumu v oblasti sociálních sítích
- 60. léta 20. stol. James C. Mitchel (antropologie) – studium sociálních systémů, formální popis příbuzenských vztahů pomocí algebry
- 70. léta 20. stol. Vzestup sociologie – Social Networks Journal, Stanley Milgram (experiment Malý svět, 6 stupňů odloučení), Mark Granovetter (slabé a silné vazby)
- 80. léta 20. stol. Rozvoj výpočetní techniky, postupné rozšiřování počítačových sítí
- 90. léta 20. stol Masivní rozvoj počítačových sítí, především Internetu, vznik programů pro analýzu sociálních sítí

Uplatnění analýzy sociálních sítí

- Pop-kultura: hry, TV, noviny, magazíny, . . .
- Obchodní sféra: nové nástroje pro konzultanty, znalostní management, . . .
- Akademické prostředí: mnoho různých oborů (lingvistika, politologie, sociologie, . . .)

Co je sociální síť?

Formální definice: Množina uzlů a hran, kde:

- uzly jsou dvojího typu: jednotlivci a skupiny,
- hrany jsou binární relace mezi jednotlivými uzly.



Sociální síť je sociální struktura tvořená jednotlivci (nebo organizacemi) zvané „uzly“, které jsou provázány (propojeny) jedním nebo více typy vzájemných závislostí, jakými jsou přátelství, příbuzenské vztahy, společné zájmy, finanční transakce, nenávisť, sexuální vztahy, vyznání, znalosti nebo prestiž.

Uzly v sociálních sítích

- **Dva druhy:**
 - **Jednotlivci:** osoby, zvířata
 - **Skupiny:** organizace, oddělení, týmy, země, města, druhy
- **Atributy uzlu** – popisují vlastnosti postoje, názory a chování, mohou se měnit v čase (např. příjem, vzdělání, zaměstnání)
- **Analýza atributů** – soubor metod, pomocí kterých lze určit a popsat jednotlivé atributy

Relace mezi uzly

Příbuzenské je matkou, je ženou

Dle rolí je šéfem, je učitelem, je přítelem

Kognitivní ví že, je si vědom

Pocitové má rád, věří

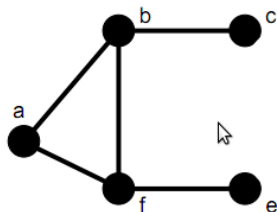
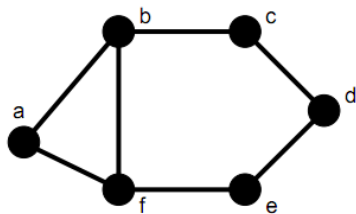
Interakční radí, mluví s, bojuje proti

Příslušnostní patří do stejného klubu, je součástí

Každá kategorie relací má vlastní strukturu a dopady na členy sociální sítě.

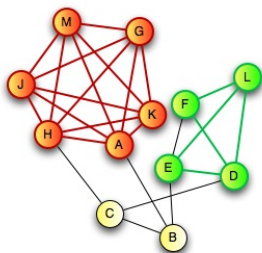
Základní pojmy (které byste měli znát)

- Komponenta souvislosti grafu
- Cesta v grafu
- Topologie grafu
- Stupeň vrcholu (vstupní/výstupní stupeň)



Klika v grafu

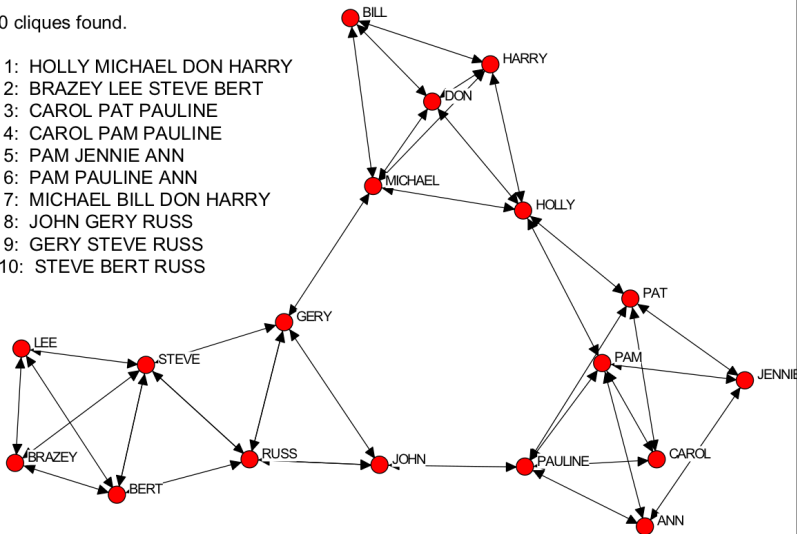
- **Klika** – největší úplný podgraf grafu, jehož vrcholy jsou spojeny hranou se všemi zbylými



Kličky – příklad

10 cliques found.

- 1: HOLLY MICHAEL DON HARRY
- 2: BRAZEY LEE STEVE BERT
- 3: CAROL PAT PAULINE
- 4: CAROL PAM PAULINE
- 5: PAM JENNIE ANN
- 6: PAM PAULINE ANN
- 7: MICHAEL BILL DON HARRY
- 8: JOHN GERY RUSS
- 9: GERY STEVE RUSS
- 10: STEVE BERT RUSS

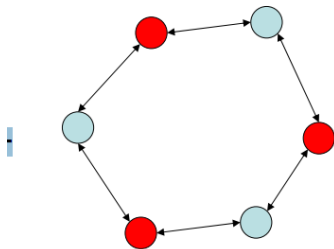
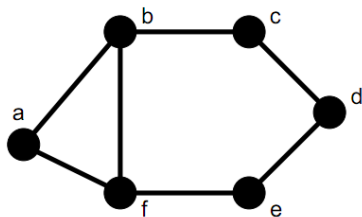


N-klika

- Definice *kliky* je pro některé případy příliš silná (skupina, kde nejsou všichni přátelé všech)
- N-klika je podgrafem grafu, ve kterém cesta mezi dvěma uzly není větší než N (pro $N = 2$: přátelé mých přátel)
- 1-klika = klika
- Pro sociology výzkum klik, kde $N > 2$ skýtá mnoho zásadních výzev

Problémy N-klik

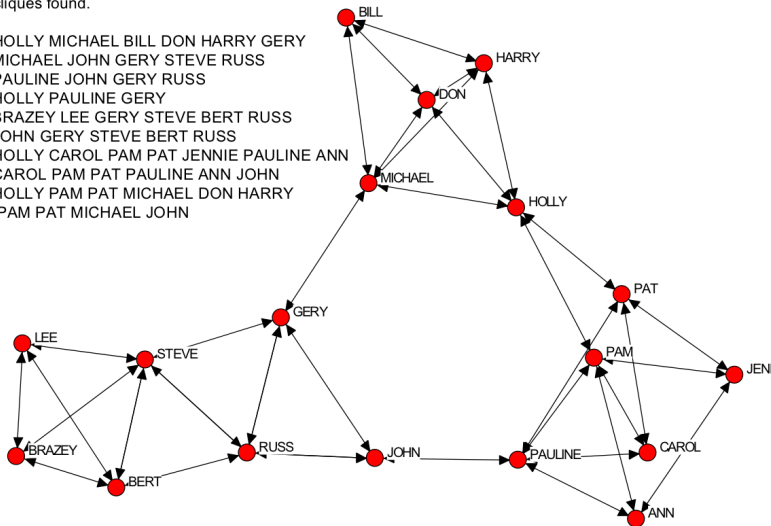
- Překryvy: $\{a,b,c,f,e\}$ a $\{b,c,d,f,e\}$ jsou obě 2-kliky
- Příslušnost do kliky je splnitelná pro nečleny
- I 2-kliky mohou být nekompaktní (cesta přes uzly mimo kliku): červené vrcholy patří do stejné 2-kliky, ačkoliv spolu nesousedí



2-kličky – příklad

10 2-cliques found.

- 1: HOLLY MICHAEL BILL DON HARRY GERY
- 2: MICHAEL JOHN GERY STEVE RUSS
- 3: PAULINE JOHN GERY RUSS
- 4: HOLLY PAULINE GERY
- 5: BRAZEY LEE GERY STEVE BERT RUSS
- 6: JOHN GERY STEVE BERT RUSS
- 7: HOLLY CAROL PAM PAT JENNIE PAULINE ANN
- 8: CAROL PAM PAT PAULINE ANN JOHN
- 9: HOLLY PAM PAT MICHAEL DON HARRY
- 10: PAM PAT MICHAEL JOHN



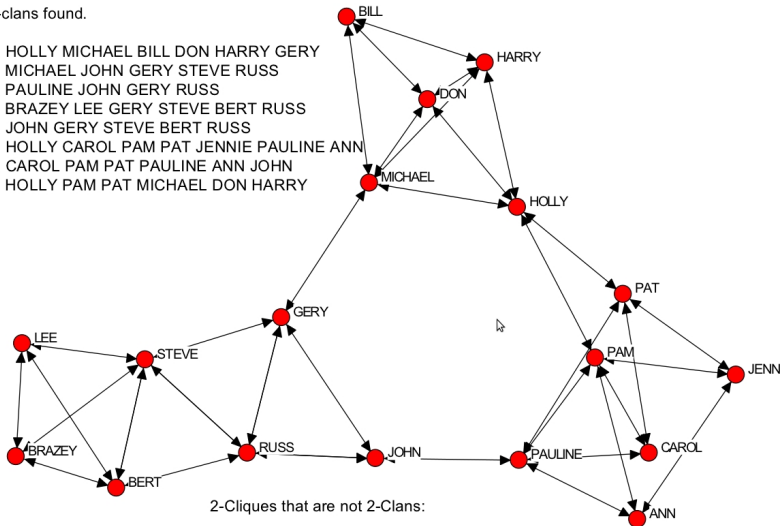
N-klan

- Podobné N-klice, cesta však nesmí vést přes uzly, které nejsou součástí N-klanu
- Příslušnost uzlů do N-klanu je dána cestami dlouhými nejvýše N hran bez použití uzlů mimo N-klan
- Vznikají tak více kompaktní podgrafy než v případě N-klik
- Problémy N-klanů:
 - Je $\{a,b,c,f\}$ 2-klan? Ukažte všechny 2-klany.
 - Je jich mnohem méně a nemusí tak zcela reflektovat realitu.
 - Různé N-klany se mohou překrývat.

2-klany – příklad

8 2-clans found.

- 1: HOLLY MICHAEL BILL DON HARRY GERY
- 2: MICHAEL JOHN GERY STEVE RUSS
- 3: PAULINE JOHN GERY RUSS
- 5: BRAZEY LEE GERY STEVE BERT RUSS
- 6: JOHN GERY STEVE BERT RUSS
- 7: HOLLY CAROL PAM PAT JENNIE PAULINE ANN
- 8: CAROL PAM PAT PAULINE ANN JOHN
- 9: HOLLY PAM PAT MICHAEL DON HARRY



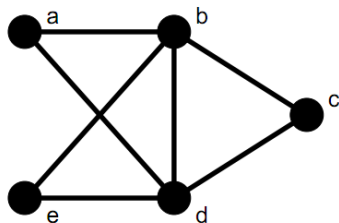
2-Cliques that are not 2-Clans:

- 4: HOLLY PAULINE GERY
- 10: PAM PAT MICHAEL JOHN

K-plex

- Množina uzlů, ve které všechny uzly sousedí s ostatními s výjimkou nejvýše K uzlů
- 1-plex = 1-klika
- Je-li $K = 2$, všechny uzly v 2-plexu jsou spojeny s alespoň $N - 2$ dalšími uzly
- K-plexu je zpravidla velmi mnoho a překrývají se

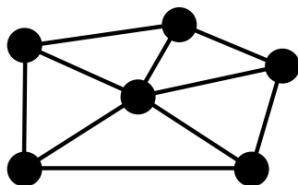
K-plex – příklad



Is $\{a, b, d, e\}$ a 2-plex?

Is $\{a, b, c, d, e\}$ a 2-plex?

Is $\{a, b, d\}$ a 2-plex?



Is the graph as a whole a 2-plex?

Is it a 3-plex?

Základní pojmy analýzy sociálních sítí I.

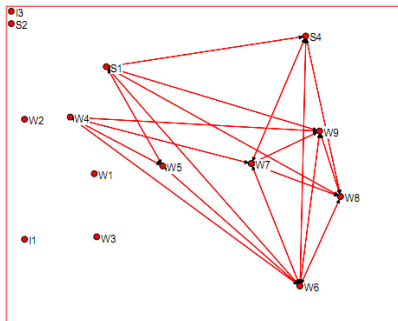
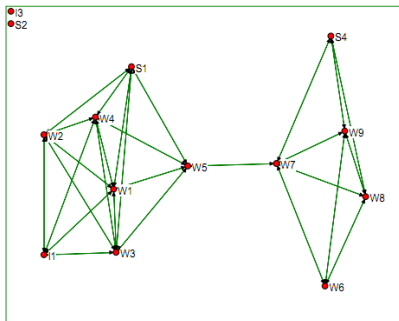
- Zabývá se zkoumáním vztahů a vzorů chování v rámci komunity
- Vztahy jsou spletité a obvykle mají více rovin (viz kategorie relací)
- Zaměřuje se primárně na analýzu **relací** mezi aktéry, nikoliv na aktéry samotné



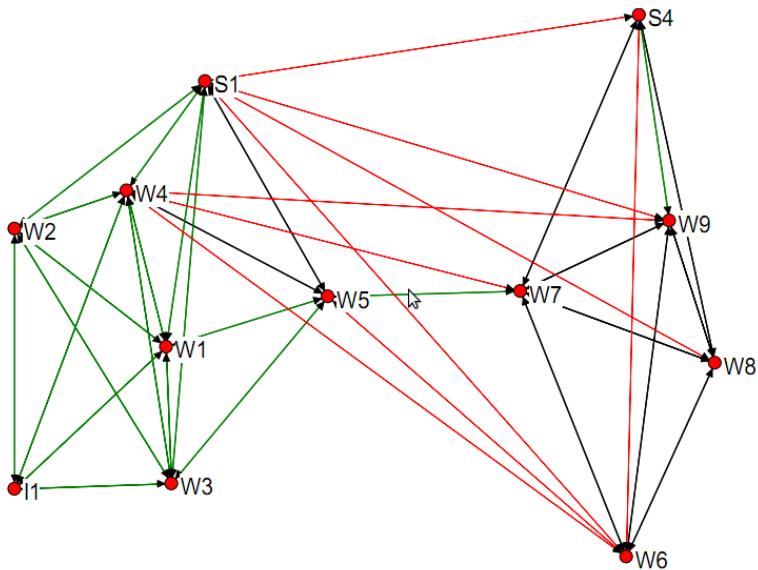
Základní pojmy analýzy sociálních sítí II.

- Relační data** – kontakty, vazby na ostatní, informace o skupinách aktérů, které definují spojení (relaci) mezi dvěma aktéry. Relační data nemohou být omezeny pouze na vlastnosti jednotlivce (atributy).
- Analýza sítě** – soubor metod, které se používají pro zkoumání a popis relací mezi jednotlivými uživateli

Příklady sítí I.



Příklady sítí II.



Získávání dat pro analýzu

- Dříve se data zapisovala ručně do tabulek, které byly více či méně efektivně organizovány, dnes se používají relační databáze
- Atributy – data uložena v **matici dat**, osoba = řádek, ve sloupcích jsou atributy
- Relace – data uložena v **matici spojení** popisujících spojení mezi uzly

Standard Data Matrix

		Variables			
		1	2	...	m
Cases	1				
	2				
	3				
	⋮				
	n				

Transposed Data Matrix

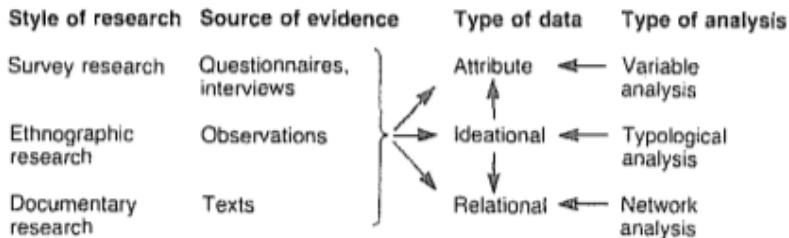
		Cases			
		1	2	...	n
Variables	1				
	2				
	3				
	⋮				
	m				

Affiliations

		Affiliations		
		A	B	C
Cases	1	1	0	0
	2	1	0	0
	3	1	0	0



Metody získávání dat



Sběr dat dnes

- Kolekce dat z webu & automatické zpracování:
 - snazší jak pro respondenty, tak pro zpracovatele.
- Pasivní sběr elektronických dat:
 - záznamy telefonních rozhovorů
 - zpracování znalostí, jež je těžké předat (znalost jazyků)
 - web cookies
 - ...

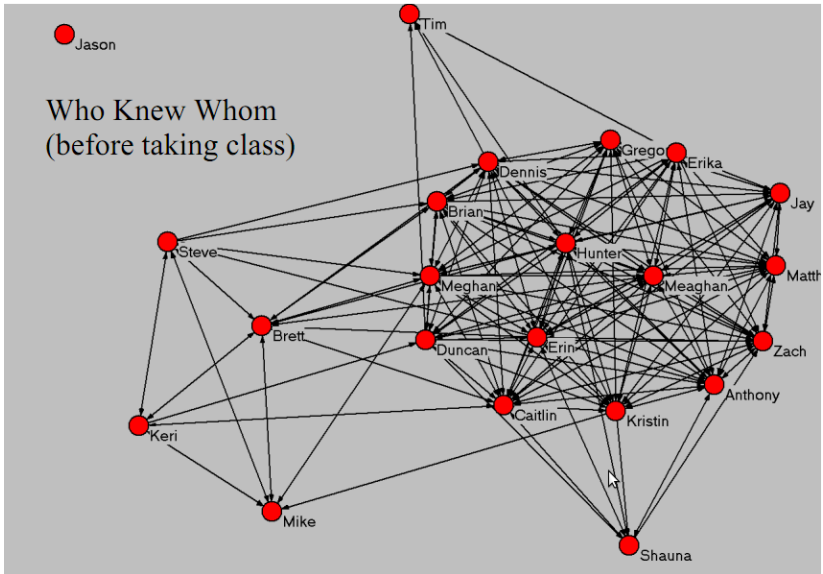
Sociocentrický vs. egocentrický pohled na síť

- **Sociocentrický pohled** se zaměřuje na vlastnosti sociální sítě jako celku. Pozice jednotlivých aktérů je charakterizována relativně vůči ostatním.
- **Egocentrický pohled** se zaměřuje na jednotlivé aktéry. Zkoumá privátní sítě jednotlivců, jejich vazby na okolí a jejich strukturu.

Tabulka pohledů a aktérů

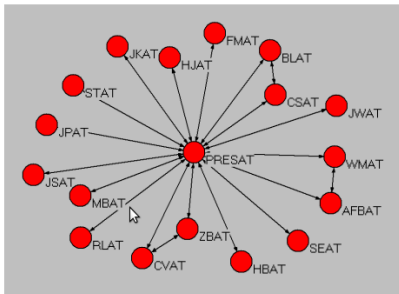
	Sociocentrický pohled	Egocentrický pohled
1 typ aktérů	Vazby mezi všemi aktéry stejného typu	Vazby mezi množinou aktérů, kteří jsou přímo spojeni s jedním konkrétní osobou
2 typy aktérů	Vazby mezi všemi aktéry dvou různých typů	Vazby mezi dvěma různými množinami aktérů, kteří jsou přímo spojeni s konkrétní osobou

1 typ aktérů – sociocentrický pohled



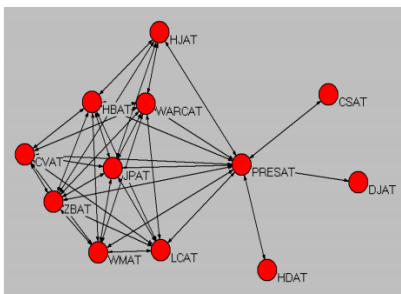
1 typ aktérů – egocentrický pohled

Carter Administration meetings



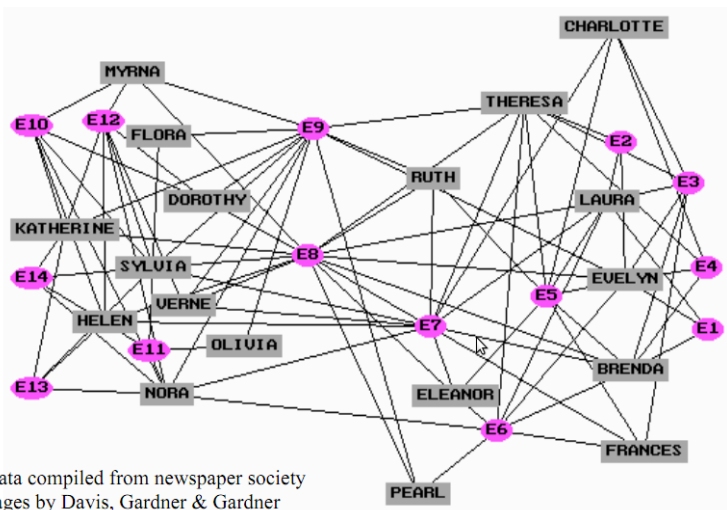
Year 1

Data courtesy of Michael Link

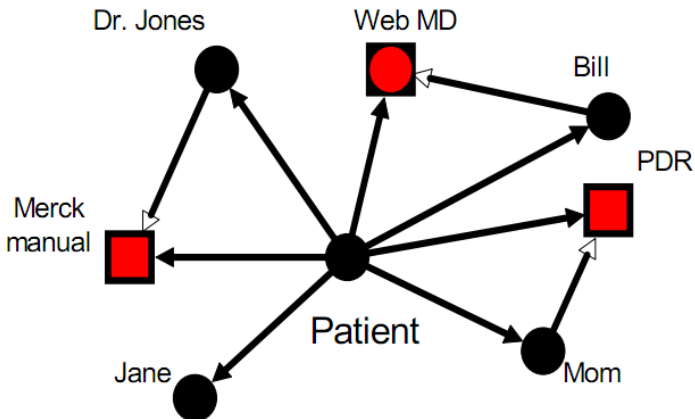


Year 4

2 typy aktérů – sociocentrický pohled



2 typy aktérů – egocentrický pohled



Kompletnost a hustota grafu

- **Kompletnost** grafu =
($\langle \text{počet uzlů grafu} \rangle - \langle \text{počet izolovaných uzlů} \rangle$)
/ $\langle \text{počet uzlů grafu} \rangle$
- Příklad: Graf o 20 vrcholech s 5 izolovanými uzly má kompletnost = 0.75
- Čím více je graf kompletní, tím vyšší je jeho hustota

- Sociocentrická hustota grafu o 91 uzlech, kde pouze 16 je propojeno hranami je 18 %
- Egocentrická hustota se liší pro každé dva vrcholy

$$\text{Hustota}(\text{neorient.}) = \frac{1}{n * (n - 1) / 2}$$

$$\text{Hustota}(\text{orient.}) = \frac{1}{n * (n - 1)}$$

Centralita I.

- Lokální centralita
 - definována, jako stupeň vrcholu grafu (počet hran vedoucích z daného uzlu (neorient. graf), resp. počet vstupních/výstupních hran (orient. graf))
 - Lokální centralita nemůže být kritériem pro srovnávání dvou různě velkých grafů
- Relativní lokální centralita
 - poměr skutečného počtu hran ku maximálnímu počtu hran
 - Př.: uzel stupně 25 v grafu o 100 vrcholech má relativní lokální centralitu 0,25; uzel stejného stupně v grafu o 30 vrcholech má relativní lokální centralitu 0,86.
 - lze ji použít i pro srovnání různě velkých grafů, nelze ji použít pro grafy různých kategorií relací

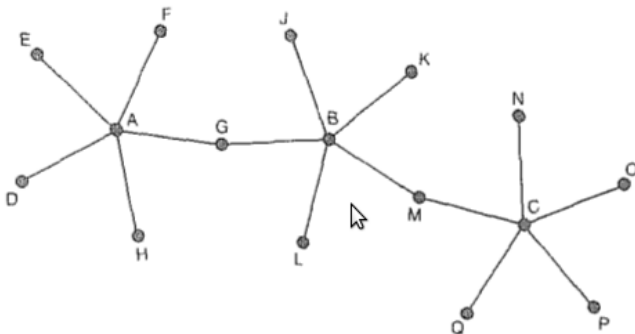
Centralita II.

- Globální centralita (blížkost)
 - Freeman (1979, 1990) navrhl metodu měření **globální centrality** založenou na pojmu „blížkosti“ uzlů
 - Globální centralita je vyjádřena jako vzdálenost (délka cesty) mezi různými vrcholy grafu.
 - Vrchol je globálně centrální, je-li suma délek všech jeho cest k většině ostatních uzlů nejmenší (má nejkratší cesty k většině ostatních uzlů)
- Užitečnost globální centrality
 - udává očekávaný čas doručení libovolné informace přenášené po síti (např. Síť drbů – centrální aktér slyší drby jako první)

Stupeň centrality

- Udává míru vystavení informací procházející grafem (např. Síť drbů – centrální aktér bude mít vždy přesnější informace než aktéři na odlehlých koncích grafu)
- Může být interpretován jako příležitost k ovlivňování & přímá ovlivnitelnost
- Předpovídá různé situace: např. šíření nemocí, uspokojení z práce

Globální centralita a stupeň centrality



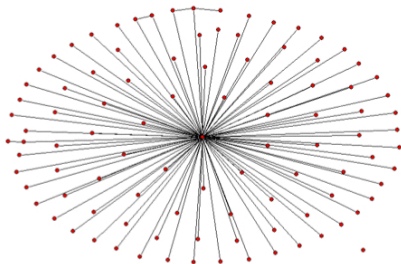
		A, C	B	G, M	J, K, L	All other points
Local centrality	Absolute	5	5	2	1	1
	Relative	0.33	0.33	0.13	0.07	0.07
Global centrality		43	33	37	48	57

Dosah centrality

- Množina vrcholů, které jsou vzájemně dosažitelné v nejvýše k krocích. Pro $k = 1$ = stupeň centrality.
- 2krokový dosah říká analytikovi, jaký je poměr lidí v síti, kteří mohou být dosaženi libovolným jiným člověkem přímo, nebo prostřednictvím jednoho prostředníka (přátelé mých přátel)
- Jedná se o dobrý ukazatel schopnosti člověka získat nějaké zdroje nebo informace od ostatních

Centralizace

- Hustota popisuje obecnou úroveň koheze v grafu
- Centralizace udává rozsah koheze, která je soustředěna kolem jednoho či více ohnisek
- Definice: Rozdíl mezi lokální centralitou každého uzlu a globálně centrálním uzlem



Obecné hypotézy

- Příslušníci stejné skupiny mají podobné výstupy: názory, chování, postoje, ... (atributy)
- Příslušnost ke skupině tak může predikovat některé atributy jedince
- Díky komunikaci mezi členy skupiny dochází k jejich vzájemnému ovlivňování

Základní informace

- Mapování sociálních vztahů mezi učiteli a zaměstnanci školy
- Výzkum probíhal na střední škole v Albertě v Kanadě (556 studentů)
- **Cíle studie**
 - identifikovat klíčové aktéry sociálních vztahů
 - systematicky zmapovat sociální vazby a následně je analyzovat
 - nalézt metodu vhodnou pro zachycení sociální struktury učitelů a dalších zaměstnanců na střední škole

Penelope Hawe and Laura Ghali Use of social network analysis to map the social relationships of staff and teachers at school Health Educ. Res. (2008) 23(1): 62-69 first published online February 7, 2007 doi:10.1093/her/cyl16

Sběr dat

- Mapováno pět kategorií relací, které byly zvoleny na základě počáteční diskuse a pilotního testu:
 - Zním dotyčného/dotyčnou jménem
 - Zním dotyčného/dotyčnou osobně
 - Pravidelně s dotyčným/dotyčnou hovořím
 - Ptám se dotyčného/dotyčné na záležitosti týkající se školy
 - Setkávám se s dotyčným/dotyčnou i mimo školu
- Použit dotazník rozdaný respondentům
- Otázky zaměřené na běžné operace a rutinní vztahy
- Zúčastnilo se 50 osob (94 %), 28 žen, 22 mužů
- 30 učitelů, 18 dalších zaměstnanců, ředitel a jeho zástupce

Vstupní předpoklady

- Výzkum byl navržen tak, aby byly jasně vymezeny jednotlivé grafy
- Ne všechny vztahy jsou symetrické (např. učitel A může znát učitele B jménem, ale naopak to platit nemusí)
- Za symetrické byly považovány: „setkávání mimo školu“ a „pravidelné konverzace“
- V ostatních případech byly relace symetrické pouze pokud to obě účastnické strany deklarovaly explicitně
- Každá z pěti kategorií tvoří vlastní síť

Metody analýzy

- Zjištění stupně centralizace sítě
- Míra 2krokové dosažitelnosti (schopnost získat informace nebo zdroje přímo a pomocí prostředníka)
- Pozice jednotlivců v rámci sítě
- Vliv pracovního zařazení: učitelé, administrativní pracovníci, ostatní personál
- Vliv pohlaví

Velikost egocentrické sítě a 2kroková dosažitelnost

	Velikost egocentrické sítě	2kroková dosažitelnost (%)
Setkávání mimo školu Ředitel / Zástupce	2 / 10	27 / 47
Žádost o radu Ředitel / Zástupce	37 / 35	85 / 83
Pravidelné konverzace Ředitel / Zástupce	29 / 27	75 / 73
Osobně se zná s Ředitel / Zástupce	33 / 35	98 / 98
Zná jej/ji jménem Ředitel / Zástupce	43 / 42	100 / 100

Závěry

- Ve všech pěti kategoriích je podpůrný personál na okraji sítě, v centru jsou učitelé
- Většina žen z podpůrného personálu byla na samém okraji sítě
- Hustota sítě korelovala s intenzitou vzájemných vazeb
- Hustota byla vyšší pro více povrchní vazby (znám dotyčnou osobu jménem) než pro ty společenské
- Hustota kategorie „znám dotyčnou osobu jménem“ byla nižší (65 %), než se předpokládalo
- Nikdo nebyl izolovaný, a to včetně 10 nových zaměstnanců (všichni měli vazbu alespoň na jednu osobu)
- Centrem sítě „dotazů“ byl ředitel a jeho zástupce, zároveň 7 osob bylo zcela mimo tuto síť

Shrnutí I.

- Analýza sociálních sítí je vhodná pro zkoumání: příbuzenských vztahů, strukturu skupiny lidí, řídicích formálních a neformálních procesů
- Využívá se informací o uzlech a relacích mezi nimi
- Různé druhy relačních vztahů produkují unikátní vrstvy sítě.
- Sociální sítě lze zkoumat z různých pohledů: sociocentrický, egocentrický, 1-mode, 2-mode.
- Hustota je poměr počtu hran mezi uzly ku maximálnímu počtu hran v (úplném) grafu. Pokud se všichni znají navzájem, hustota je 100 %.

Shrnutí II.

- Centralizace sítě udává míru dominance jednoho či více aktérů na síti. Může tak ukázat na klíčové osoby v síti.
- Kliky, N-kliky a N-klany jsou užitečné pro identifikaci skupin v síti
- Existuje velké množství metod pro analýzu sociálních sítí. Konkrétní volba závisí na záměru konkrétního výzkumu
- SW nástroje pro analýzu sociálních sítí (např. UCINET, Pajek) jsou nezbytné pro automatizovanou kolekci především u velkých datových sad

Shrnutí III.

- Další příklady uplatnění analýzy sociálních sítí:
 - Hledání klik nebo spojitých komponent podgrafu, volba závisí na cílech a účelu, které jsou stanoveny
 - Identifikace osob se strategickým významem, identifikace přirozených vůdců a pomocníků v dané komunitě

Doporučená literatura

- John P. Scott. Social Network Analysis: A Handbook. Sage Publications Ltd; 2nd edition (March 2000)
- John P Scott (Editor), Peter Carrington (Editor). The SAGE Handbook of Social Network Analysis. Sage Publications Ltd (May 25, 2011)
- Penelope Hawe and Laura Ghali Use of social network analysis to map the social relationships of staff and teachers at school Health Educ. Res. (2008) 23(1): 62-69. 2007.