Úvod do analýzy sociálních sítí

Ego síť a e-mailové sítě

2024-25

Ego network

- *Ego síť* se tvoří kolem jednoho uzlu sítě – ega.
- Součástí ego sítě je *ego* a všichni jeho sousedé, dohromady tvoří tzv. *sousedství* (neighborhood).
  - Někdy se pracuje nejen se sousedy, ale se všemi vrcholy, které mají nějakou předem definovanou maximální vzdálenost N od ega (N step neighborhood).
- Součástí ego sítě jsou všechny hrany mezi vrcholy patřící do sousedství.
- Ego síť může být vážená nebo nevážená.
  - Váha hrany ve vážené síti reprezentuje intenzitu interakce mezi vrcholy.
  - Pokud je síť nevážená, pak hrana mezi vrcholy reprezentuje zvolenou míru minimální interakce mezi vrcholy

Co nebo kdo může být egem

- Pojem ego souvisí s lidskými (ale např. i se zvířecími) společenstvími a jeho sousedé se označují jako alters.

- Obecněji ale za ego síť můžeme považovat síť s jedním vrcholem v centru a jeho sousedy, přičemž vrcholy sítě jsou entity stejného typu.

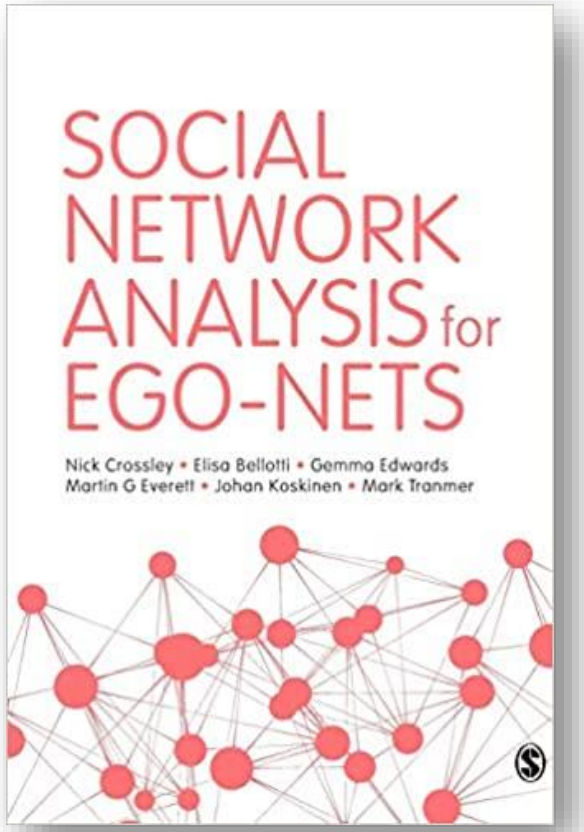
  - Může se jednat o sítě osob, skupin, organizací, společností apod.

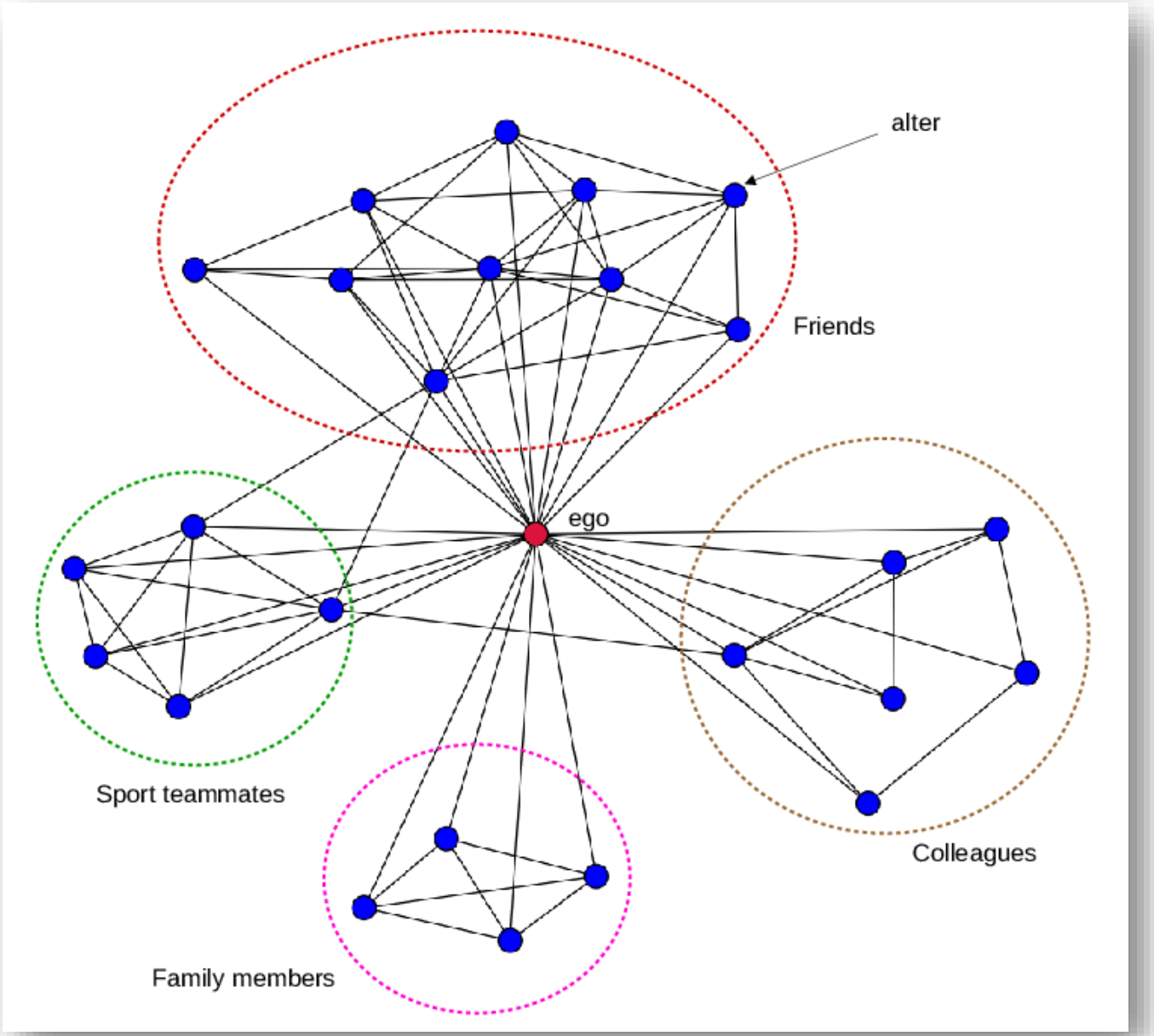
- Obecněji tedy intenzita interakce může znamenat různé věci související s různými událostmi (výměna informací, spolupráce apod.).

Jak získat ego síť

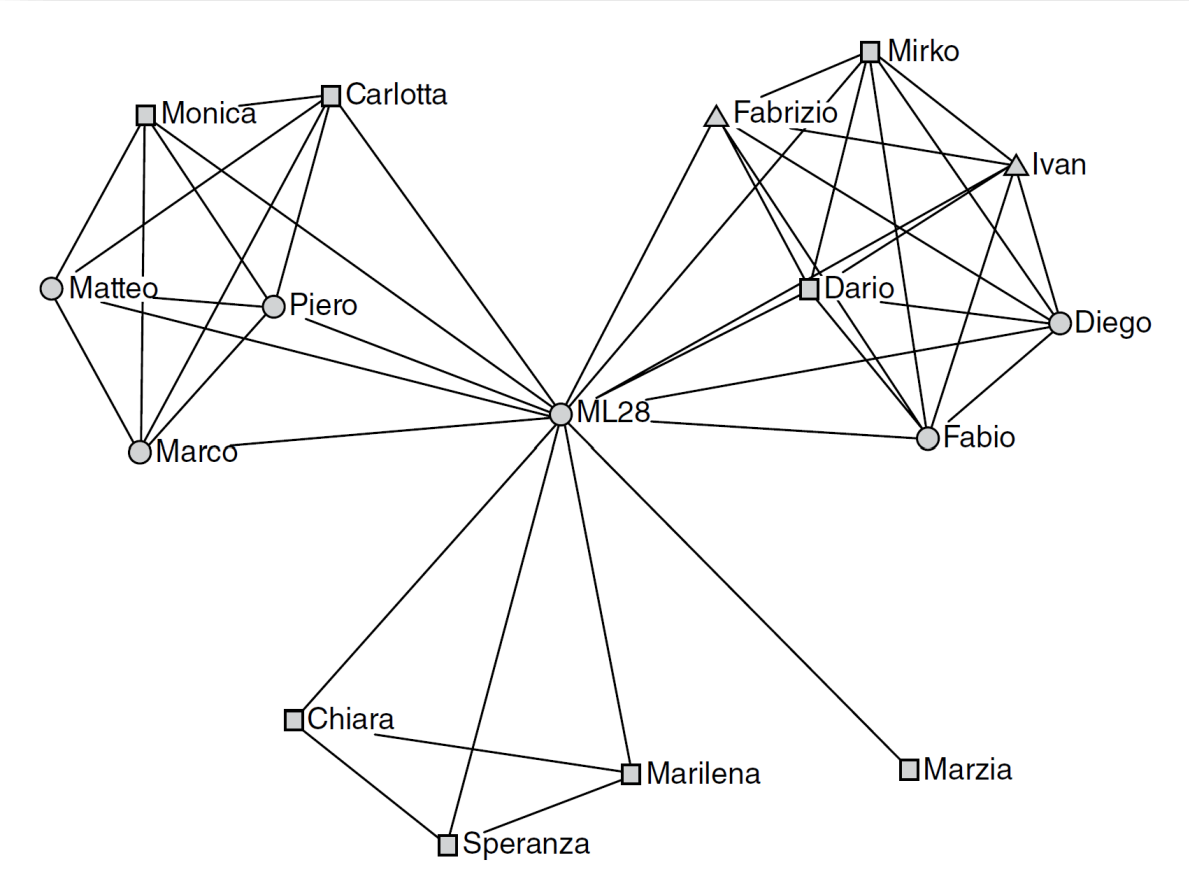
- Z libovolné sítě můžeme extrahovat tolik ego sítí, kolik je celkem vrcholů.
- Ego sítě jsou sítě jako každé jiné. Proto můžeme měřit jejich vlastnosti jako:
  - hustota (počet existujících hran vůči počtu teoreticky možného počtu hran),
  - průměrný stupeň (průměrný počet sousedů v ego síti),
  - průměrná délka nejkratší cesty mezi vrcholy,
  - komunitní struktury a komunity.

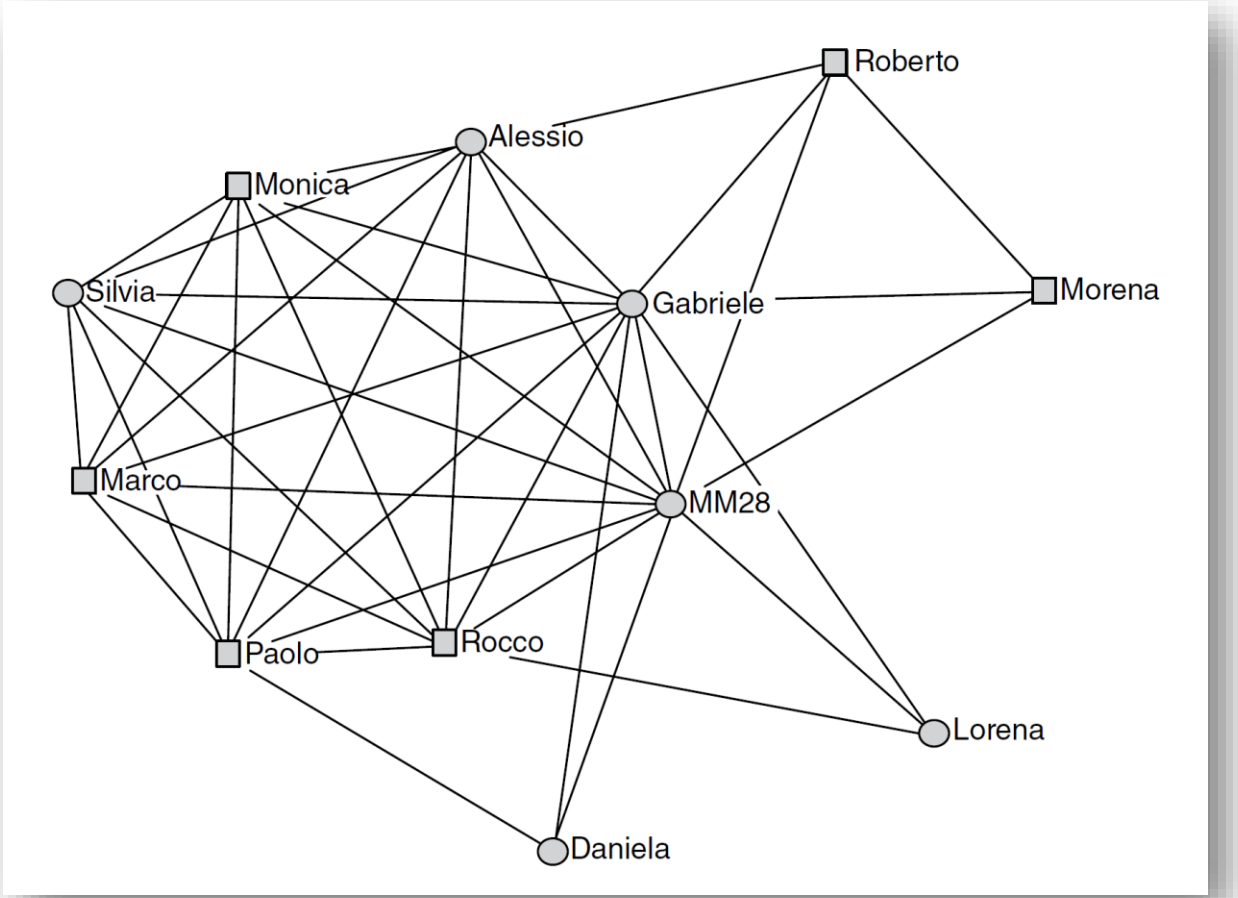
Příklad ego sítě





Sociální kapitál





E-mailová ego síť

- E-mailovou síť můžeme chápat jako ego síť v případě, že zpracováváme mailovou komunikaci z jednoho účtu.
- Egem je v tomto případě vlastník účtu, ostatní vrcholy sítě jsou ti, kterým vlastník mail odeslal, nebo od kterých mail obdržel.
- Několik emailových ego sítí (reprezentujících několik různých účtů) lze sloučit do jedné, tzv. komunikační, sítě.

Zehnalova, S., Horak, Z., Kudelka, M. (2015). *Email conversation network analysis: Work groups and teams in organizations*. In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 1262-1268).

Data pro emailovou ego síť

- Musíme najít způsob, jak extrahovat maily z mailové schránky (např. gmail, MS Exchange apod).
- Dostaneme seznam mailů v nějaké čitelné struktuře.
- Je potřeba provést čištění, tzn. ponechat k dalšímu zpracování pouze ty maily, které přinášejí informační hodnotu, jako např.
  - Vyřadit např. všechny maily, ve kterých není uveden vlastník účtu ani jako odesílatel ani mezi příjemci.
  - Vyřadit maily, jejichž odesílatelem je někdo z předem připraveného blacklistu odesílatelů nebo domén.
  - Vyřadit maily, které obsahují v předmětu slova z předem připraveného blacklistu.
  - …apod.

Co můžeme extrahovat z mailů?

- Ve vyčištěných mailových zprávách lze provádět:
  - Seřadit je podle času odeslání.
  - Vytvořit seznam všech emailových adres (odesílatel, příjemci).
  - Vytvořit seznam všech domén, ze kterých nebo kam byly maily odeslány.
  - Vytvořit seznam všech slov, která se vyskytují např. v předmětu zprávy.
  - Vytvořit skupinu mailů označenou jako *konverzaci* začínající jedním mailem a rozvíjejících se podle toho jak bylo na maily v této konverzaci odpovídáno a jak byly přeposílány.

A jak tedy vytvořit emailovou ego síť?

- Ego, tedy centrální vrchol ego sítě je vlastník účtu.
- Ostatní emailové adresy uvedené jako odesílatelé nebo příjemci budou další vrcholy sítě.
- Hrana mezi vrcholy bude existovat v případě, že jsou emailové adresy, které tyto vrcholy reprezentující, uvedeny v jednom mailu jako odesílatel nebo příjemci.
- Jeden mail se tedy promítá do tzv. kliky sítě (úplný podgraf).

Jaké problémy budeme řešit?

- Síť konstruovaná ze schránky s průměrně intenzivní několikaletou komunikací bude velmi hustá a s klikami, které mají mnoho (desítky) vrcholů.

- Některé hrany budou reprezentovat jeden mail, jiné i několik stovek mailů.

- Velké množství vrcholů může mít vysoký stupeň (mnoho sousedů).

Jaké máme metody řešení?

- Můžeme zpracovat maily pouze z malého časového období (1 rok/měsíc).
- Můžeme si dopředu připravit seznam odesílatelů a vzít v úvahu pouze maily, kde odesílatelem byl někdo z tohoto seznamu.
- Můžeme si dopředu připravit seznam domén a vzít úvahu pouze maily, kde je odesílatelem někdo z domény v tomto seznamu.
- Můžeme si dopředu sestavit seznam slov z předmětu a vzít v úvahu pouze maily, které mají v předmětu alespoň jedno ze slov v tomto seznamu.
- Můžeme vytvořit emailovou konverzaci popsanou dříve a v mailech této konverzace pracovat jen s odesílateli.
- …apod.

Strukturální řešení

- Předpokládejme, že budeme pracovat s váhami hran. Pak váha (síla) hrany bude odpovídat počtu mailů, ve kterých je dvojice vrcholů (emailových adres) společně.

- Pak můžeme hrany seřadit od nejvyšší po nejnižší váhu a hrany s nízkou váhou můžeme odstranit.

- Poté můžeme odstranit vrcholy, kterým nezůstala žádná hrana.

Prořídnutí: Co je nízká váha?

- Můžeme například předpokládat, že nechceme příliš hustou síť, což bude např. odpovídat námi očekávanému průměrnému stupni <k> = 10.
- Pokud máme po konstrukci sítě např. 100 vrcholů a 3000 hran, pak platí <k> = 2 * 3000 / 100, což je 60. My tedy potřebujeme, aby platilo <k> = 2 * 500 / 100.
- Po uspořádání hran podle váhy pak ponecháme v síti pouze prvních 500 hran.
- Vrcholy, kterým nezbyla žádná hrana (outliers), pak můžeme ze sítě odstranit. Tím ale nemusí zůstat v síti 100 vrcholů a stupeň se může o něco zvýšit.

Příklad

Firemní komunikace

- Záznam komunikace zaměstnanců jedné firmy mezi sebou, se zákazníky a uživateli za ~5 let.
- Všechna data jsou anonymizována – jména, doména, slova z předmětu mailu (tzv. *stopslova* jsou odstraněna).
- Obsah mailu a přílohy se nezpracovávaly.

```
</Message><Message Sender="name00008@domain0004.cz" Sent="2010-03-15T14:40:28.000Z"
MessageId="f75bf701-ecd5-48ac-a8ed-b865a0c60230" InReplyTo="0a376a9a-80b9-48c0-9ad3-
fb2c5db61186">
  <Recipient Type="To">name00188@domain0029.cz</Recipient>
  <Recipient Type="To">name00188@domain0029.cz</Recipient>
  <Recipient Type="To">name00189@domain0029.cz</Recipient>
  <Recipient Type="To">name00189@domain0029.cz</Recipient>
  <Recipient Type="Cc">name00140@domain0029.cz</Recipient>
  <Recipient Type="Cc">name00056@domain0002.cz</Recipient>
```

Zaměstnanci firmy

- Jeden SW vývojář se třemi mailovými adresami, další dva SW vývojáři, SW vývojář na specifických platformách, CEO, tým leader + architekt.

```
<Accounts>
  <Account smtpAddress="name00001@domain0001.com" />
  <Account smtpAddress="name00002@domain0002.cz" />
  <Account smtpAddress="name00003@domain0003.cz" />
  <Account smtpAddress="name00004@domain0004.cz" />
  <Account smtpAddress="name00005@domain0005.cz" />
  <Account smtpAddress="name00006@domain0004.cz" />
  <Account smtpAddress="name00007@domain0004.cz" />
  <Account smtpAddress="name00008@domain0004.cz" />
  <Account smtpAddress="name00009@domain0004.cz" />
  <Account smtpAddress="name00010@domain0004.cz" />
</Accounts>
```

Struktura jednoho záznamu (e-mailu)

- Sender: Adresa odesílatele
- Sent: Datum a čas odeslání
- MessageId: unique ID (může chybět)
- InReplyTo (nepovinné): MessageId předcházejícího mailu v konverzaci (může chybět)
- Subject (optional): Předmět – skupina slov
- Recipient: Adresa příjemce typu “To” nebo “Cc”

```
</Message><Message Sender="name00008@domain0004.cz" Sent="2010-01-13T17:00:48.000Z"
MessageId="9fce8b1b-2db3-447a-b422-aba74d8e4264" InReplyTo="1f29f3b6-f77e-42a5-b272-
f37adcf6b50d" Subject="term00053 term00054 term00055 ">
  <Recipient Type="To">name00045@domain0028.cz</Recipient>
  <Recipient Type="Cc">name00056@domain0002.cz</Recipient>
```

Úkoly

- Zkonstruujte síť (sítě) z dat ze cvičení a vyzkoušejte alespoň jednu z metod, které prořídnou konstruovanou síť (např. Enron).
- V rozsáhlé síti, která byla konstruována a vyčištěna z několika agregovaných anonymizovaných účtů se pokuste najít významné uživatele (Ega).
  - Slovo „významný“ si definujte např. vysokým stupněm vrcholu.
- V konstruované Ego-síti najděte komunity.