Wprowadzenie do analizy sieci społecznych

Mikołaj Morzy
Agnieszka Ławrynowicz

Instytut Informatyki
Poznań, rok akademicki 2010/2011

Początki: analiza sieci

Początki SNA sięgają zarówno nauk społecznych, jak i ogólnych metod: **analizy sieci** i **teorii grafów**.

Analiza sieci zajmuje się formułowaniem i rozwiązywaniem problemów wyrażonych za pomocą struktur sieciowych, reprezentowanych najczęściej w postaci grafów.

Teoria grafów dostarcza abstrakcyjnych metod analizy grafów. W połączeniu z technikami wizualizacji i badania sieci tworzy podstawę analizy sieci społecznych (SNA).

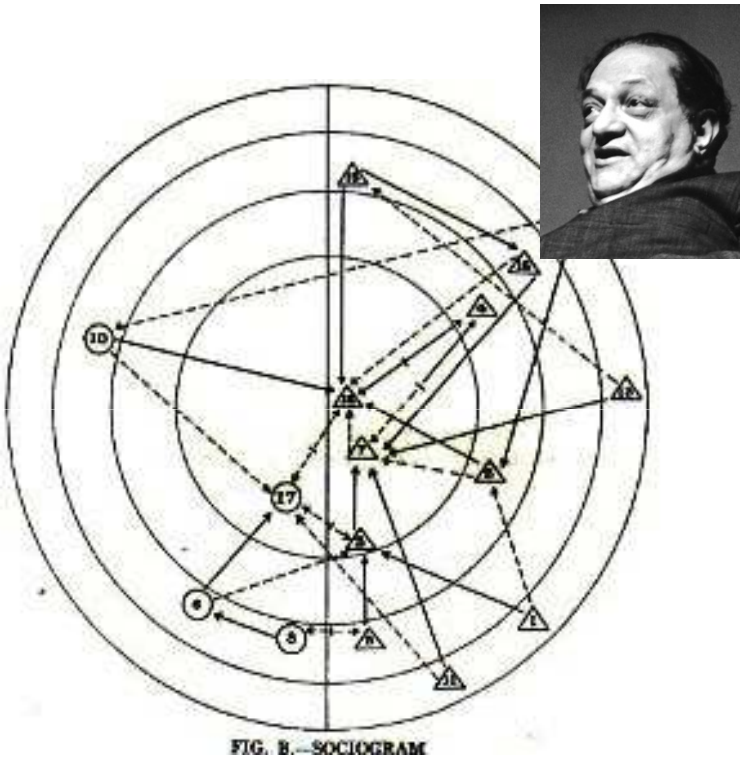
SNA to przede wszystkim specyficzna perspektywa analizy: skupienie się nie na indywidualnych jednostkach lub makrostrukturach, lecz **związkach** między jednostkami, grupami i instytucjami.

Początki: nauki społeczne

Badanie społeczeństw z perspektywy sieciowej oznacza postrzeganie jednostek jako zanurzonych w sieci powiązań. Wytłumaczenia zachowań społecznych poszukuje się w strukturze sieci, zamiast w cechach osobniczych. Taka perspektywa nabiera szczególnego znaczenia w "usieciowionych społeczeństwach"

Badania SNA w ramach nauk społecznych były historycznie przeprowadzane przede wszystkim przez matematyków, fizyków, informatyków, biologów (nauki o sieciach i grafach)

Koncepcja wykorzystania powiązań między jednostkami nie jest nowa, ale dopiero nowe zbiory danych i nowe metody obliczeniowe umożliwiły zastosowanie SNA na wielką skalę.



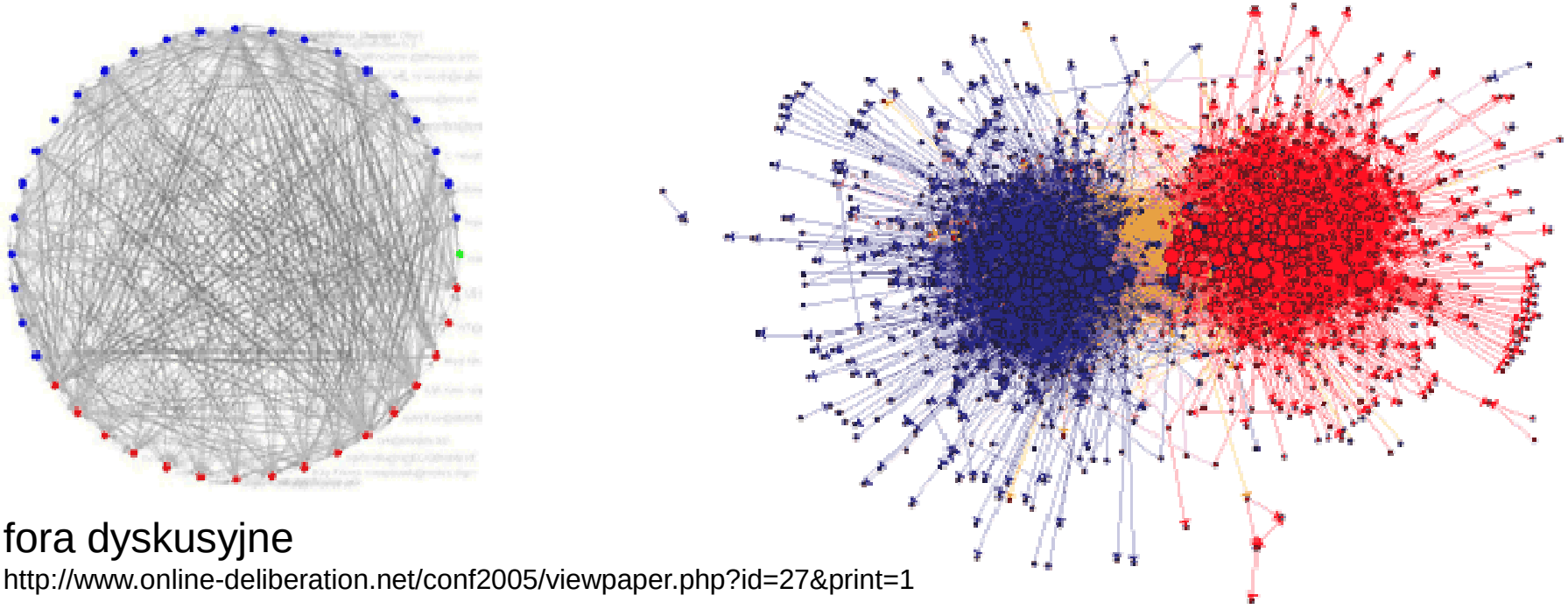
Socjogram przygotowywany przez Jacoba Moreno, genialnego amerykańskiego socjologa, który położył podwaliny pod SNA w latach 30-tych XX wieku

Wizualizacja w naukach społecznych

Porównanie wzorców dyskusji na tematy polityczne

Blogi

http://www.blogpulse.com/papers/2005/AdamicGlanceBlogWWW.pdf



fora dyskusyjne

http://www.online-deliberation.net/conf2005/viewpaper.php?id=27&print=1

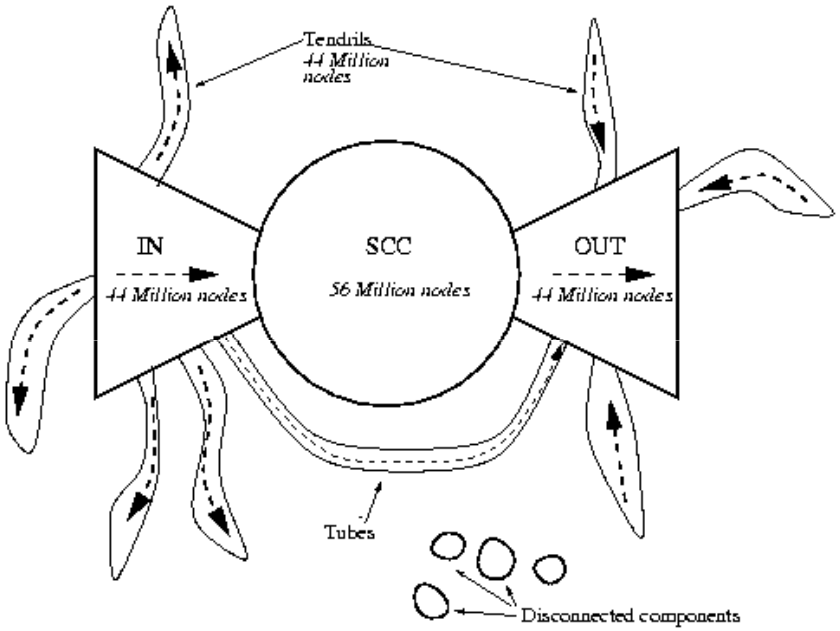
Początki: inne nauki

SNA znajduje zastosowanie przede wszystkim do badania struktur tworzonych przez ludzi, ale może być zastosowana także do badania innych zjawisk.

W informatyce SNA jest wykorzystywana do analizy przepływów informacji, badania struktur sieci WWW, itp.

W biologii SNA została zaaplikowana do badania łańcuchów pokarmowych w różnych ekosystemach i przewidywania zmian w łańcuchach w odpowiedzi na zmiany w ekosystemach.

http://www.searchengineposition.com/info/articles/bowtie.asp



Teoria muszki, czyli propozycja wyjaśnienia przyczyn popularności niektórych odnośników w Sieci
"Graph Structure of the Web", A.Broder et al., Computer Networks

SNA: zastosowania praktyczne

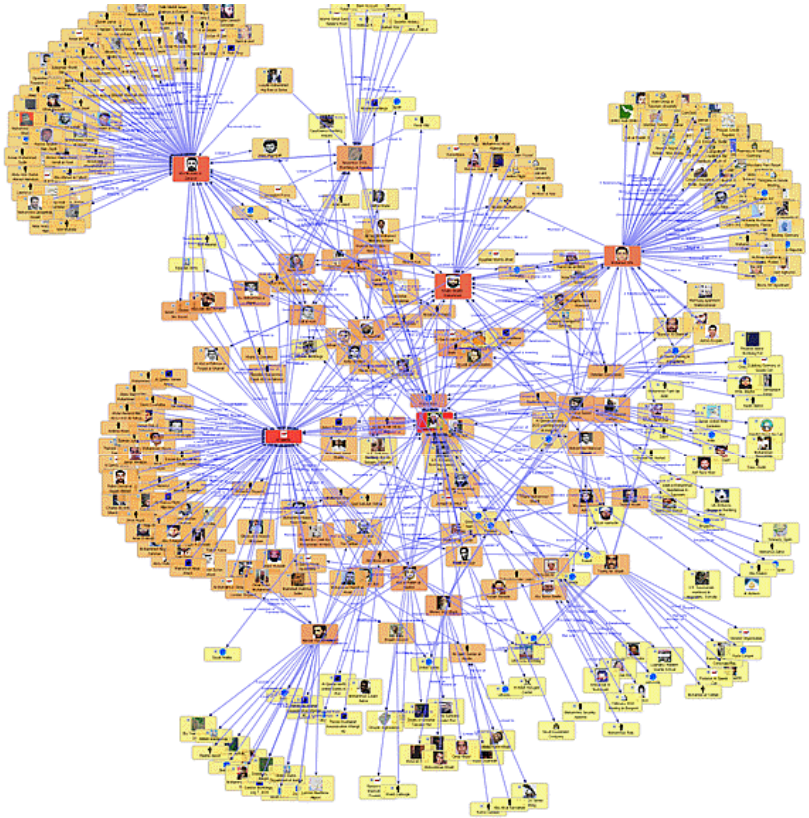
Ekonomia: SNA jest wykorzystywana do analizy i poprawy kanałów komunikacji wewnątrz organizacji

Policja: służby wywiadowcze i policja wykorzystują SNA do badania struktur siatek przestępczych i terrorystycznych

Serwisy WWW: wiele serwisów wykorzystuje SNA do znajdowania rekomendacji linków i znajomych na podstawie FoaF

Operatorzy telekomunikacyjni: korzystają z SNA do optymalizacji przepustowości i pojemności sieci

Organizacje *watchdog*: badają ukryte powiązania między administracją, przemysłem i lobbystami



Sieć powiązań między sprawcami zamachów z 11.07
Gradientowa wizualizacja ujawnia stopień bliskości między poszczególnymi aktorami.

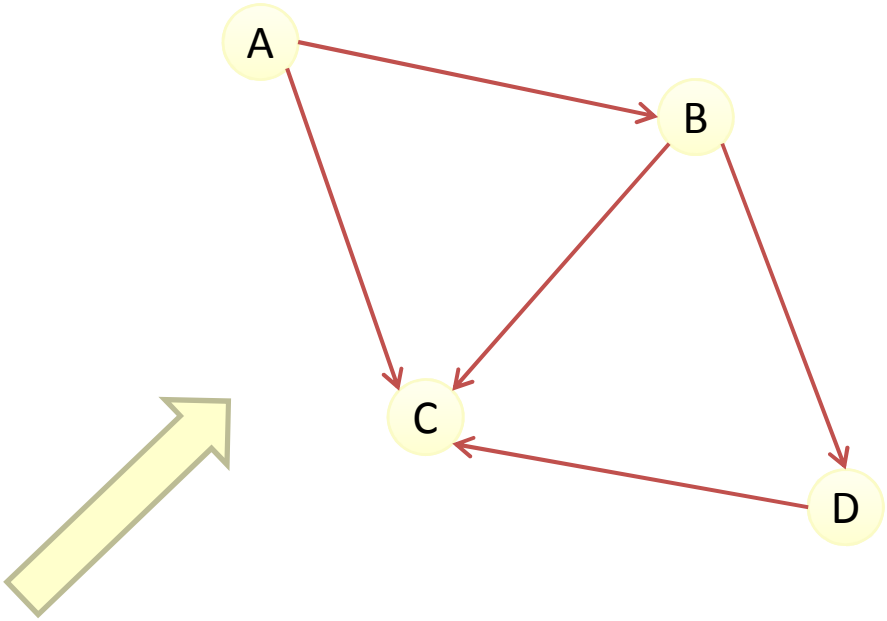
Kiedy i po co korzystać z SNA?

- Do badania dowolnej struktury sieciowej lub w trakcie optymalizacji parametrów sieci (np. przepustowości lub szybkości rozprzestrzeniania się sygnału)
- Do wizualizacji danych w celu odkrycia niewidocznych struktur i wzorców w powiązaniach i związkach między aktorami
- Do badania skuteczności dyseminacji informacji w sieci
- Do analizy ilościowej sieci
  - zachowanie i akcje poszczególnych aktorów jest często funkcją ich pozycji w sieci, a nie ról tradycyjnie przypisywanych do aktorów
  - analiza ilościowa pozwala określić role pełnione przez węzły w sieci, w tym zidentyfikować najważniejszych aktorów, na których można skupić analizę jakościową

Podstawowe pojęcia

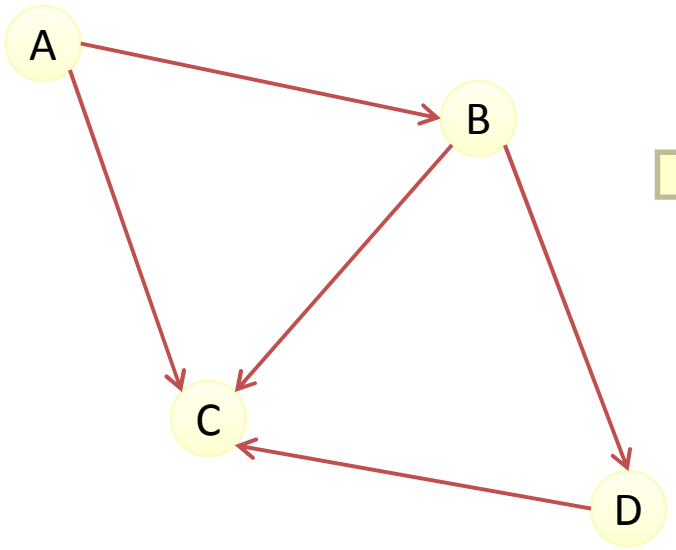
sieci	• w jaki sposób modelować i reprezentować zjawiska w postaci sieci społecznych?
powiązania	• w jaki sposób identyfikować silne i słabe punkty w sieci?
gracze	• w jaki sposób znajdować najważniejsze lub najbardziej centralne węzły w sieci?
spójność	• jak mierzyć "jakość" struktur sieciowych lub aktualnej konfiguracji?

Reprezentacja związków w postaci sieci



Adam: Bartek, powiedz Cecylii że zapraszamy ich na obiad
Bartek: Cecylia, przyjdź z mamą do nas na obiad
Bartek: Pani Doroto, czy przyjdziecie do nas na obiad?
Adam: Dorota, czy Bartek powiedział Ci o obiedzie? Koniecznie przyjdźcie!
Cecylia: Mamo, jesteśmy zaproszone na obiad.
Dorota: Dobrze Adam, na pewno wpadniemy.

Możliwe notacje grafów skierowanych



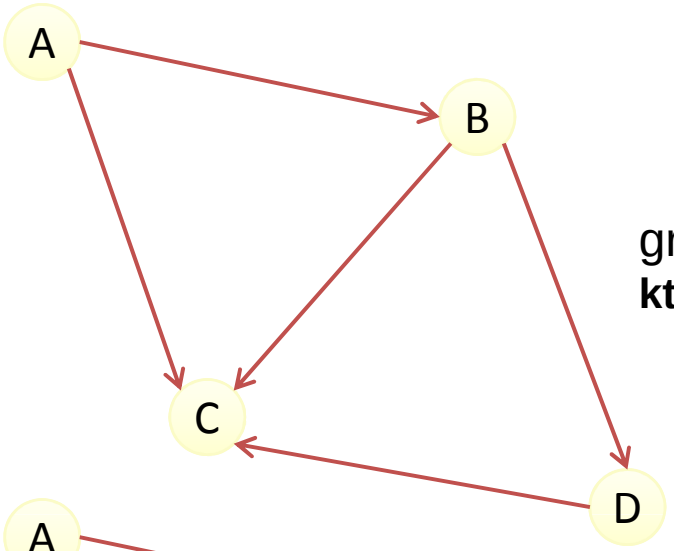
węzeł	węzeł
A	B
A	C
B	C
B	D
D	C

lista krawędzi

macierz sąsiedztwa

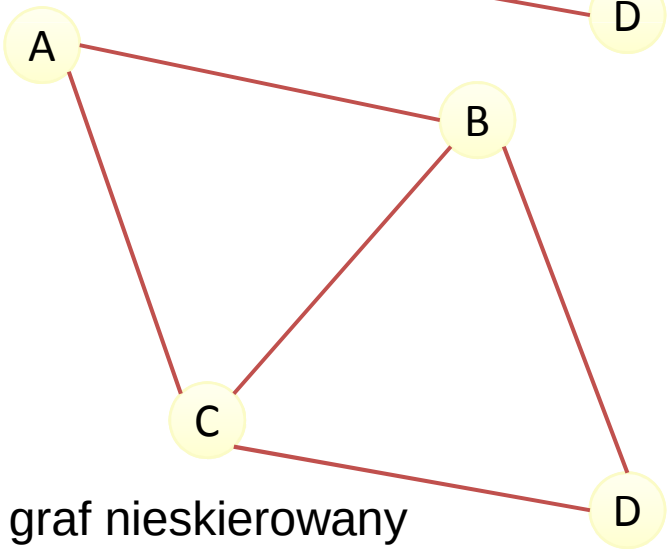
	A	B	C	D
A		1	1	0
B	0		1	1
C	0	0		0
D	0	0	1	

Możliwe notacje grafów nieskierowanych

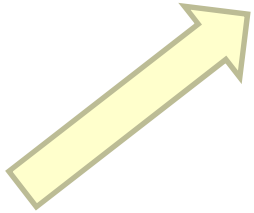


graf skierowany
kto z kim się kontaktuje?

węzeł	węzeł
A	B
A	C
B	C
B	D
D	C

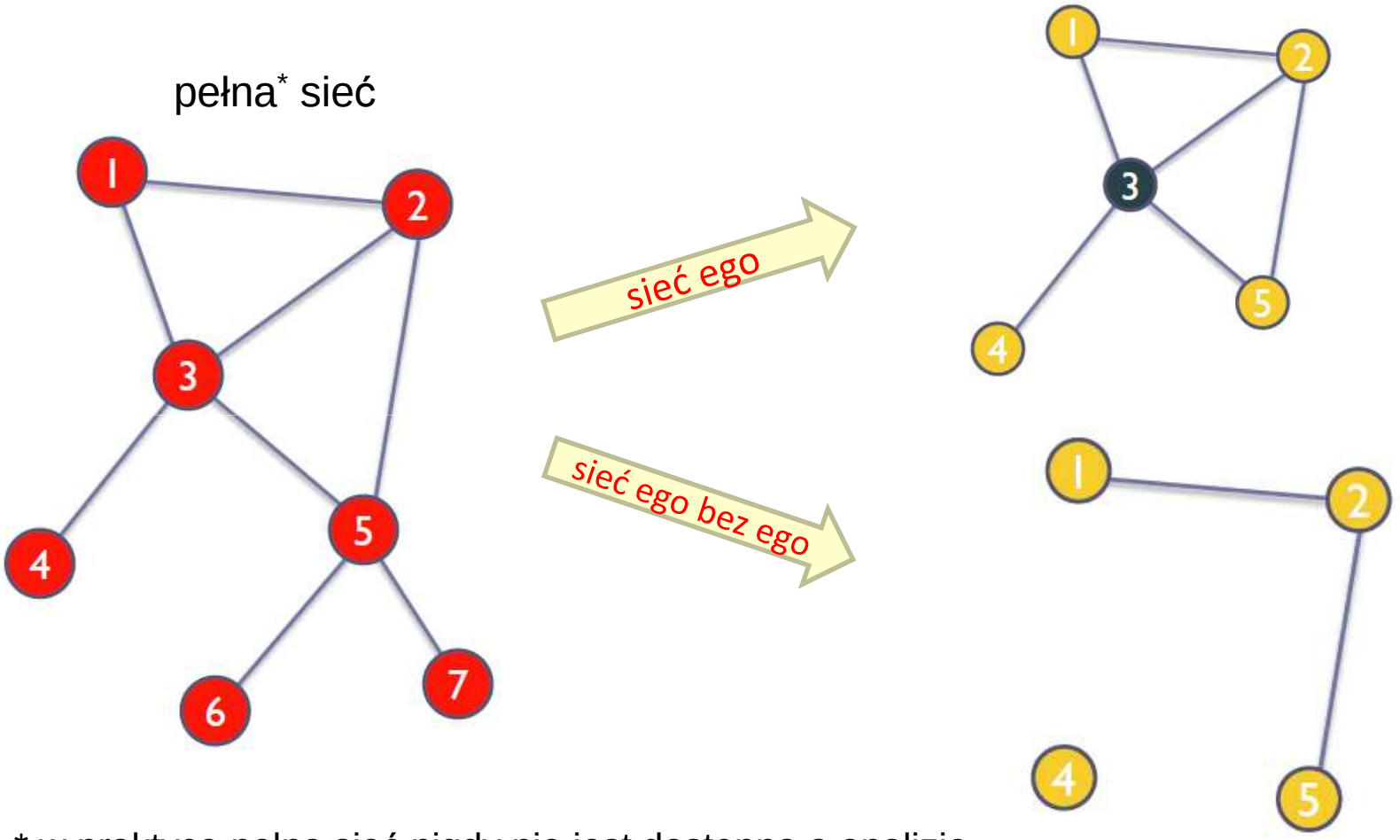


graf nieskierowany
kto kogo zna?



	A	B	C	D
A		1	1	0
B	1		1	1
C	1	1		0
D	0	1	1	

Sieci typu "ego" i sieci pełne

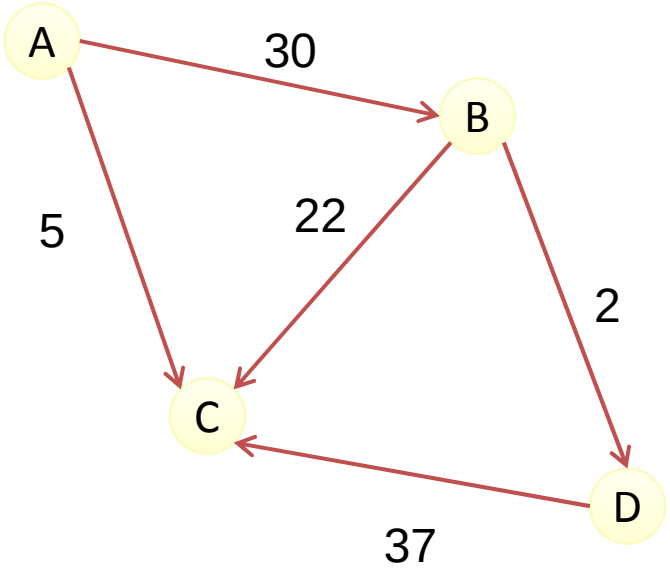


* w praktyce pełna sieć nigdy nie jest dostępna a analizie podlega wycinek sieci (problem demarkacji granic)

Podstawowe pojęcia

sieci	• w jaki sposób modelować i reprezentować zjawiska w postaci sieci społecznych?
powiązania	• w jaki sposób identyfikować silne i słabe punkty w sieci?
gracze	• w jaki sposób znajdować najważniejsze lub najbardziej centralne węzły w sieci?
spójność	• jak mierzyć "jakość" struktur sieciowych lub aktualnej konfiguracji?

Wagi krawędzi w grafach



węzeł	węzeł	waga
A	B	30
A	C	5
B	C	22
B	D	2
D	C	37

lista krawędzi

macierz sąsiedztwa

	A	B	C	D
A		30	5	0
B	30		22	2
C	5	22		37
D	0	2	37	

wagi mogą reprezentować:
- częstotliwość interakcji
- liczbę wymienionych przedmiotów
- indywidualne odczucie siły przyjaźni
- koszt komunikacji (np. odległość)
- bardziej złożone relacje

Wagi krawędzi jako siła związku

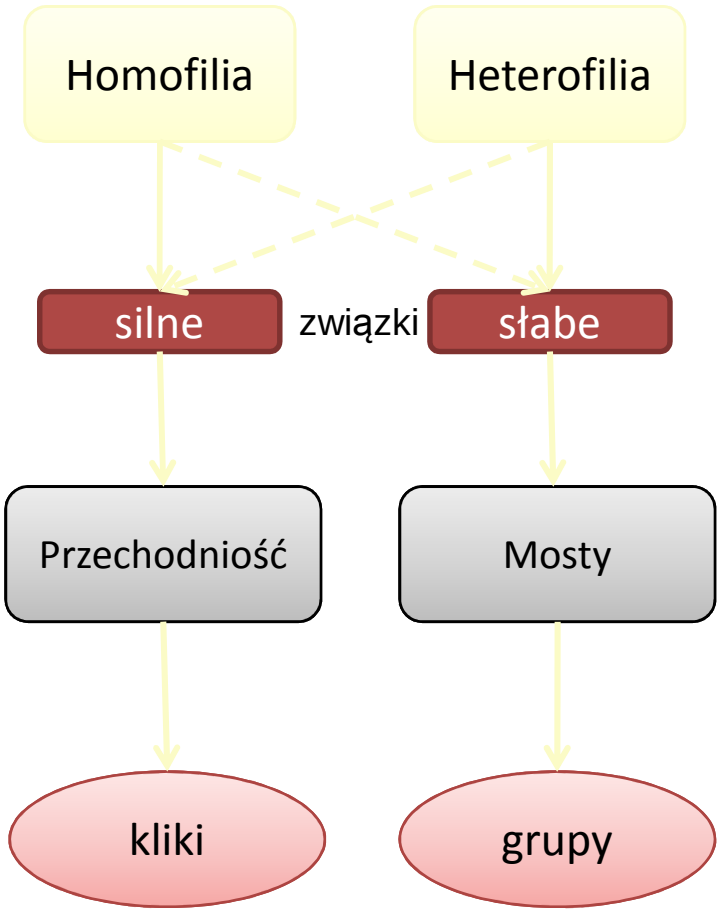
Krawędzie reprezentują interakcję, przepływ informacji i dóbr, podobieństwo, afiliację, lub związki społeczne.

Dla związków społecznych miarą siły powiązania może być:

- częstotliwość interakcji (np. komunikacji) lub wolumen przepływu (wymiana)
- wzajemność interakcji lub przepływu
- rodzaj interakcji lub przepływu
- atrybuty łączonych węzłów lub krawędzi (np. stopień pokrewieństwa)
- struktura sąsiedztwa łączonych węzłów (np. liczba wspólnych sąsiadów)

Bezpośrednie badania (np. ankiety) pozwalają lepiej określić rodzaj i siłę związku, ale miary typu *proxy* są także bardzo użyteczne.

Homofilia, przechodniość i mosty



Homofilia to tendencja do łączenia się z ludźmi o podobnych cechach (statusie, zainteresowaniach, przekonaniach, itp.)

- homofilia prowadzi do powstawania homogenicznych grup w których tworzenie związków jest łatwe
- homofilia może stanowić przeszkodę do wymiany informacji lub innowacji, heterofilia może być w wielu kontekstach pożądana

Przechodniość jest cechą powiązań, istnienie związków między A i B oraz B i C sugeruje istnienie związku między A i C

- silne związki znacznie częściej są przechodnie niż słabe związki, istnienie przechodniości jest często wyraźnym sygnałem istnienia silnych związków
- przechodniość i homofilia prowadzą do powstawania klik i pseudo-klik

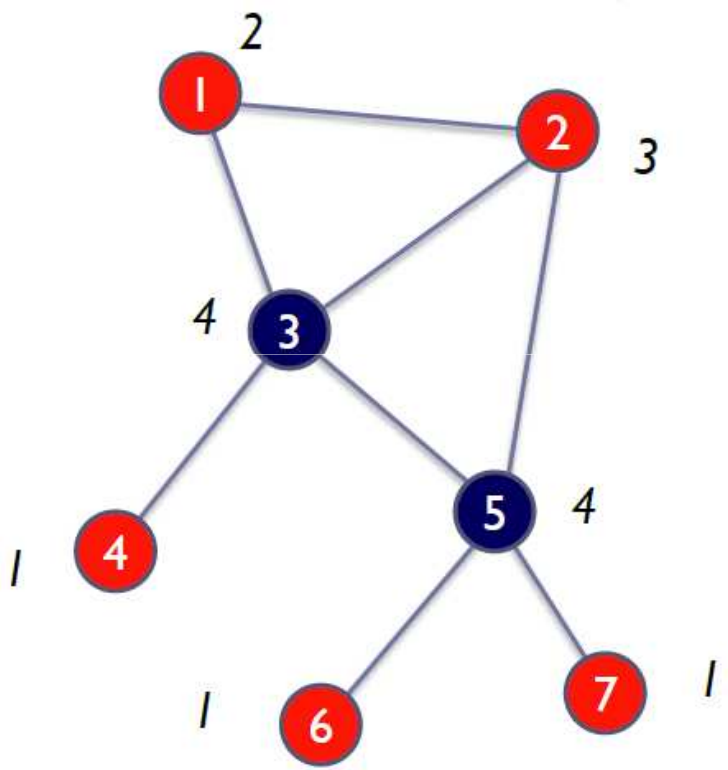
Mosty to węzły i krawędzie łączące różne grupy

- ułatwiają komunikację między grupami, zwiększają spójność, umożliwiają rozprzestrzenianie informacji i innowacji
- mostami są najczęściej słabe związki

Podstawowe pojęcia

sieci	• w jaki sposób modelować i reprezentować zjawiska w postaci sieci społecznych?
powiązania	• w jaki sposób identyfikować silne i słabe punkty w sieci?
gracze	• w jaki sposób znajdować najważniejsze lub najbardziej centralne węzły w sieci?
spójność	• jak mierzyć "jakość" struktur sieciowych lub aktualnej konfiguracji?

Stopień wierzchołka

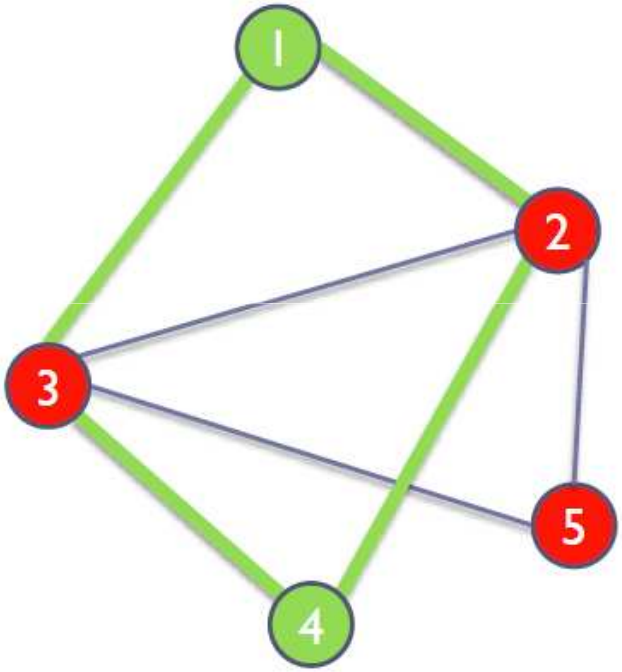


Stopniem wejściowym (ang. *in-degree*) lub wyjściowym (ang. *out-degree*) węzła nazywamy liczbę krawędzi wchodzących lub wychodzących z węzła (dla grafów nieskierowanych posługujemy się pojęciem stopnia wierzchołka)

Centralność wg stopni wierzchołków często mierzy popularność lub wpływowość węzłów

Centralność wg stopni wierzchołków jest użyteczna do określania, które węzły są kluczowe z punktu widzenia rozprzestrzeniania informacji lub wpływania na węzły położone w bezpośrednim sąsiedztwie.

Ścieżki i najkrótsze ścieżki



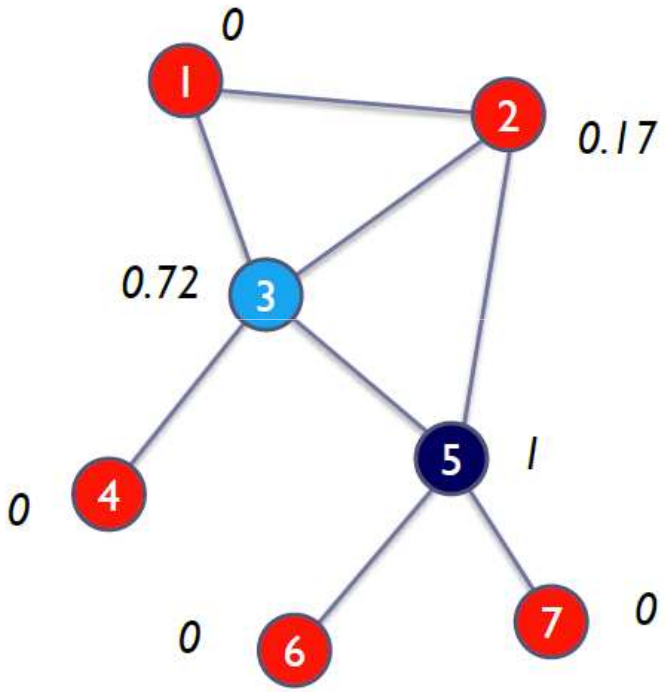
Ścieżką (ang. *path*) między dwoma węzłami nazywamy dowolną sekwencję unikalnych węzłów łączącą dane dwa węzły

Najkrótszą ścieżką między dwoma węzłami nazywamy ścieżkę o najmniejszej liczbie krawędzi. Liczbę krawędzi nazywamy odległością (ang. *distance*) między węzłami

- najkrótsze ścieżki <1,2,4> i <1,3,4>
- inne ścieżki: <1,2,5,3,4> lub <1,3,5,2,4>

Krótkie ścieżki są często pożądaną własnością sieci (np. w przypadku maksymalizacji prędkości komunikacji), istnieją też sieci, w których pożądane są długie ścieżki (np. w przypadku sieci transmisji chorób zakaźnych)

Pośrednictwo



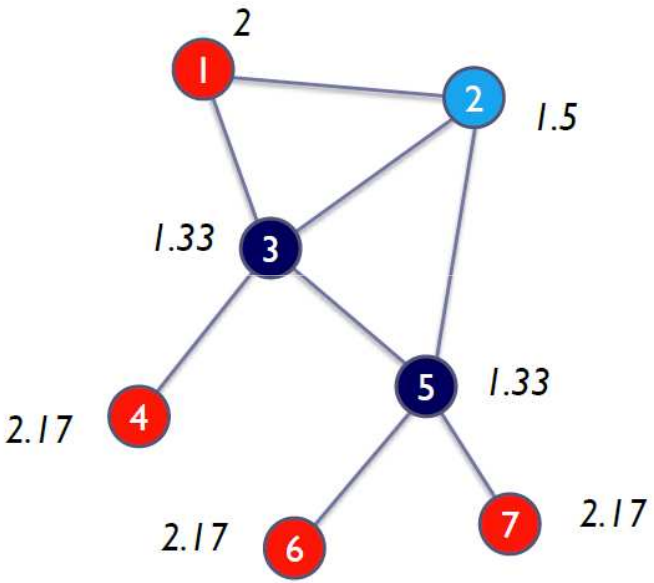
Pośrednictwem (ang. *betweeness*) węzła *v* nazywamy stosunek liczby najkrótszych ścieżek między dowolnymi dwoma węzłami przechodzących przez węzeł *v* do łącznej liczby wszystkich najkrótszych ścieżek

Czasami pośrednictwo jest normalizowane w taki sposób, aby maksymalne pośrednictwo w sieci wynosiło 1

Pośrednictwo wskazuje, które węzły są najważniejsze z punktu widzenia komunikacji między węzłami

Węzły o dużym pośrednictwie mogą być punktami utraty spójności sieci

Bliskość

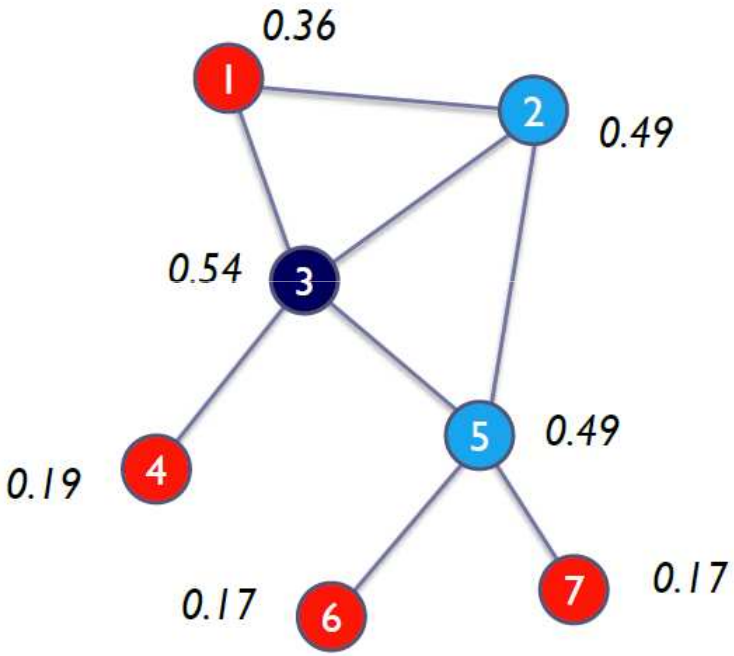


Bliskością (ang. *closeness*) węzła v nazywamy średnią długość najkrótszych ścieżek między węzłem v i wszystkimi pozostałymi węzłami (innymi słowy jest to oczekiwana odległość między węzłem v i dowolnym innym węzłem)

Bliskość jest miarą zasięgu (ang. *reach*) danego węzła, tj. miarą czasu, jaki jest potrzebny, aby z danego węzła dotrzeć do innych węzłów

- miara przydatna w sytuacji gdy podstawowym kryterium jest prędkość rozchodzenia się informacji
- niskie wartości są pożądane w przypadku gdy celem jest minimalizacja czasu propagacji informacji

Wektor własny węzła

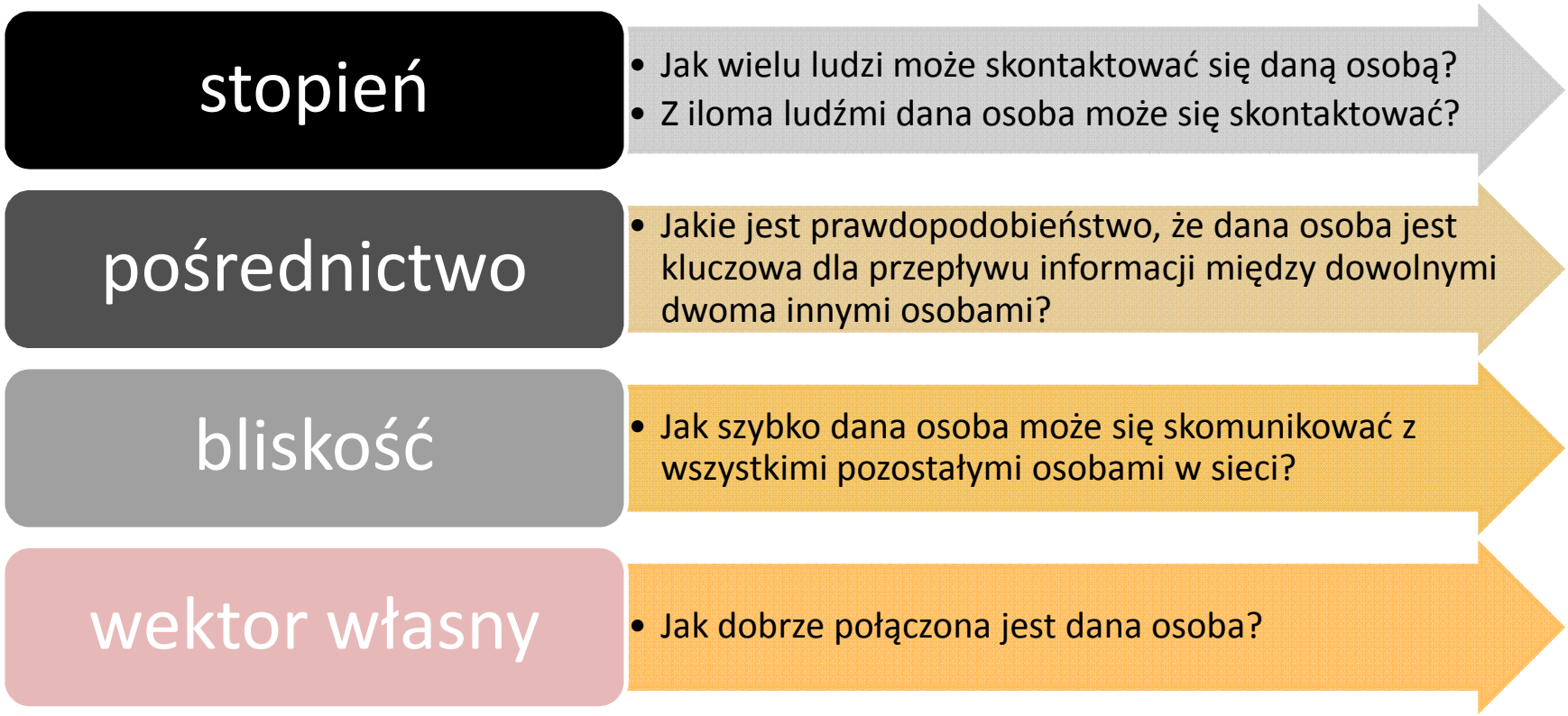


Wartość wektora własnego węzła (ang. *eigenvector*) jest proporcjonalna do wartości wektorów własnych węzłów, z którymi dany węzeł jest bezpośrednio połączony

Wektor własny węzła określa względną ważność węzła w sieci

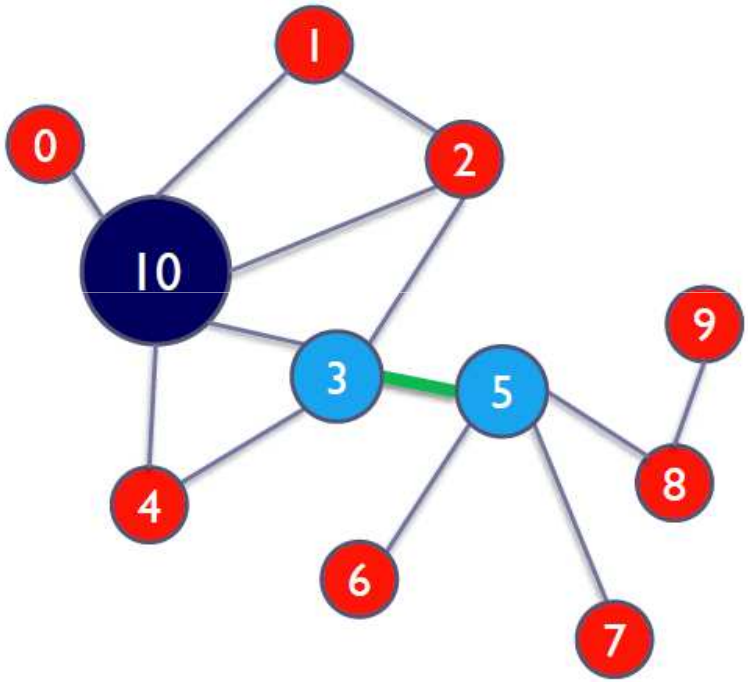
- wersją tej miary jest algorytm PageRank
- miara wektora własnego informuje, które węzły są powiązane z najbardziej powiązanymi węzłami

Interpretacja podanych miar centralności



Interpretacje praktyczne

Identyfikacja najważniejszych aktorów w sieci



W przedstawionej obok sieci węzłem o najwyższym stopniu jest węzeł 10

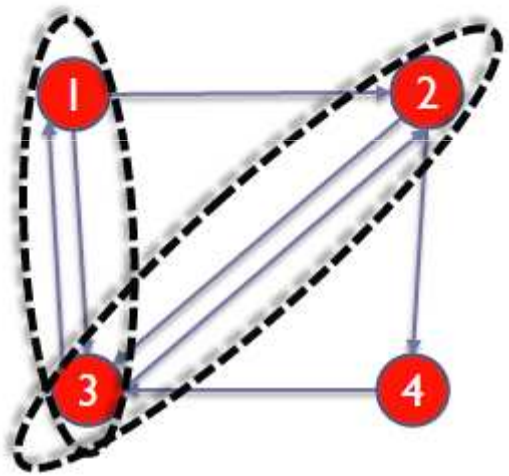
Węzły 3 i 5 mają łącznie większy stopień niż węzeł 10, dodatkowo, łącze między nimi jest krytyczne dla spójności sieci

Pod wieloma względami węzły 3 i 5 są "ważniejsze" z punktu widzenia funkcjonowania sieci niż węzeł 10

Podstawowe pojęcia

sieci	• w jaki sposób modelować i reprezentować zjawiska w postaci sieci społecznych?
powiązania	• w jaki sposób identyfikować silne i słabe punkty w sieci?
gracze	• w jaki sposób znajdować najważniejsze lub najbardziej centralne węzły w sieci?
spójność	• jak mierzyć "jakość" struktur sieciowych lub aktualnej konfiguracji?

Wzajemność



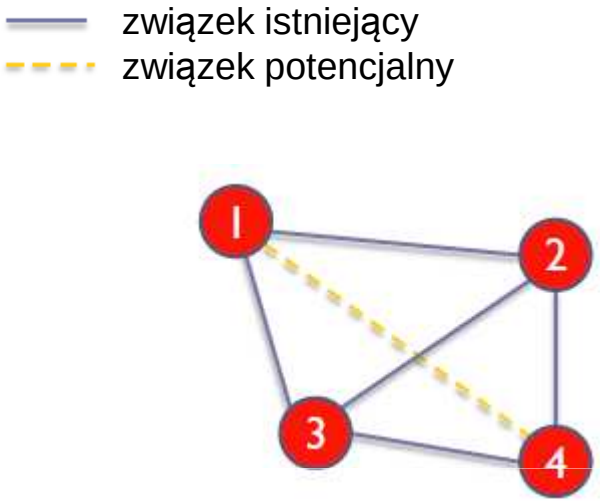
wzajemność = 0.4

Wzajemność (ang. *reciprocity*) to stosunek związków zwrotnych do liczby wszystkich związków występujących w sieci

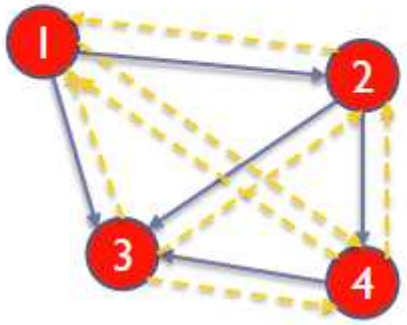
Interpretacja konkretnej wartości miary wzajemności jest silnie zależna od kontekstu analizy

Wzajemność jest użyteczną miarą obopólności wymiany informacji w sieci, która z kolei jest popularnie wykorzystywaną miarą spoistości sieci. Wzajemność jest zdefiniowana jedynie w grafach skierowanych

Gęstość



gęstość = 5/6 = 0.83



gęstość = 5/12 = 0.42

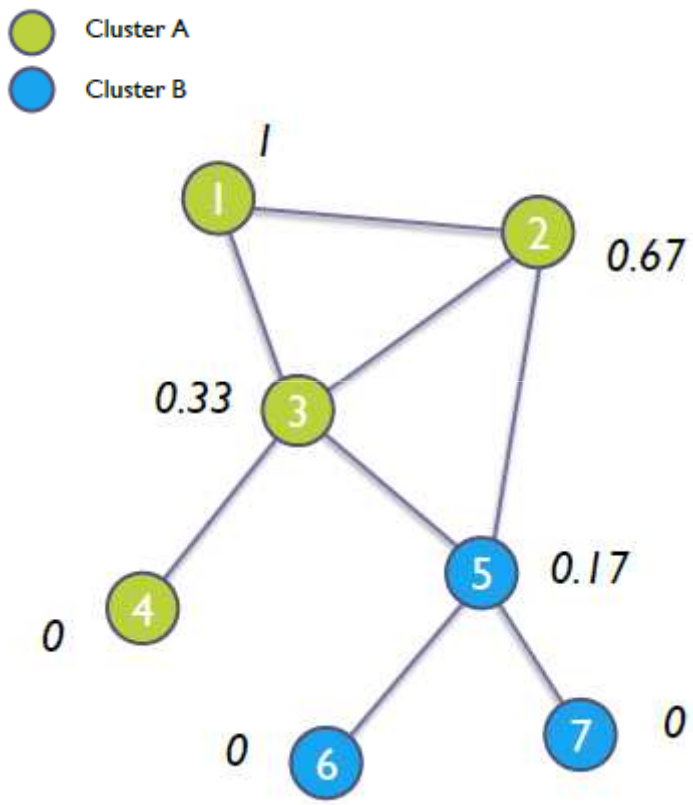
Gęstość (ang. *density*) sieci to stosunek liczby istniejących związków w sieci do liczby wszystkich potencjalnych związków w sieci (który wynosi *n*(n-1) / 2* dla liczby wierzchołków n w grafie nieskierowanym)

Gęstość to popularna miara kompletności sieci, lub stopnia usieciowienia sieci (przykładowo, klika ma gęstość 1)

Graf skierowany jest scharakteryzowany dwukrotnie mniejszą gęstością niż odpowiadający mu graf nieskierowany

Gęstość jest użyteczną miarą kontrolną gdy ta sama analiza jest przeprowadzana dla różnych obszarów sieci

Współczynnik grupowania



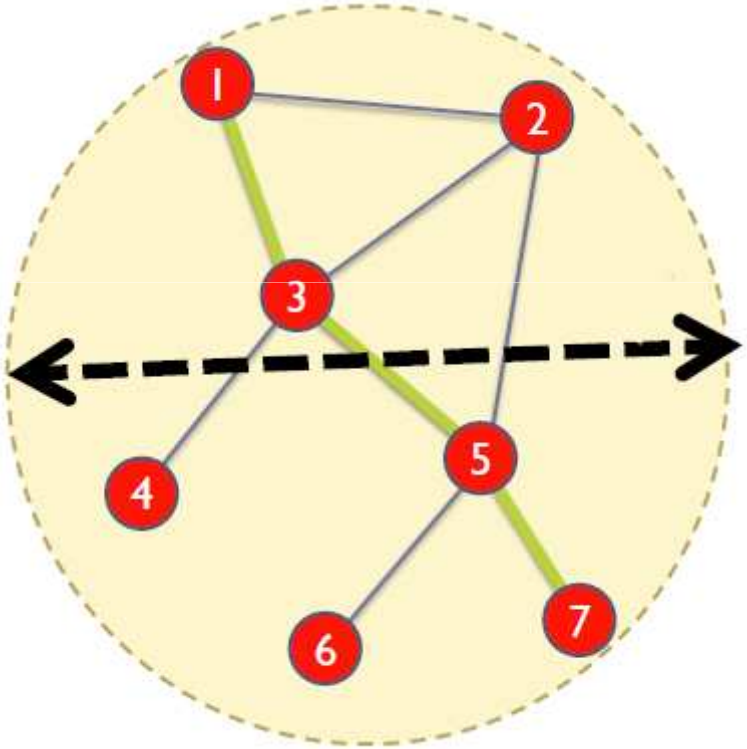
współczynnik grupowania = 0.31

Współczynnik grupowania (ang. *clustering coefficient*) dla danego węzła jest gęstością bezpośredniego sąsiedztwa danego węzła, gdzie sąsiedztwo jest zdefiniowane jako zbiór wszystkich węzłów bezpośrednio powiązanych z danym węzłem

Współczynnik grupowania dla całej sieci jest średnią współczynników grupowania wszystkich węzłów wchodzących w skład sieci

Algorytmy grupowania próbują maksymalizować współczynnik grupowania sieci. Celem algorytmów grupowania jest odkrywanie wspólnot i społeczności w ramach sieci

Średnica, największa i najmniejsza odległość



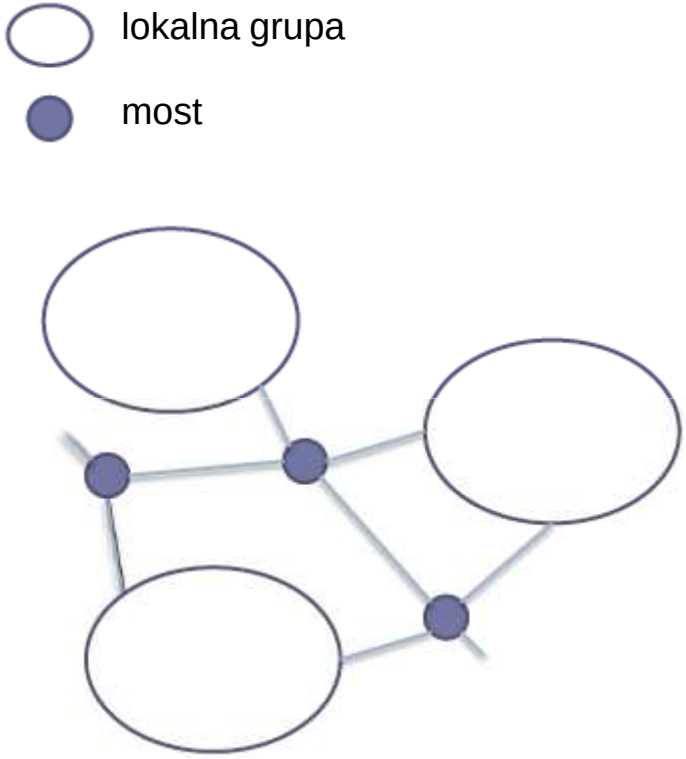
Średnicą sieci (ang. *diameter*) nazywamy długość najdłuższej spośród najkrótszych ścieżek łączących dowolne dwa węzły w sieci

Średnica sieci jest często bardziej znaczącym parametrem (szczególnie jeśli chodzi o propagację informacji w sieci) niż liczba węzłów lub liczba krawędzi

- rzadkie sieci charakteryzują się za zwyczaj większą średnicą
- sieci bezskalowe posiadają relatywnie małe średnice

Alternatywą dla średnicy sieci jest średnia odległość między węzłami (średnia najkrótszych ścieżek między wszystkimi parami węzłów)

Zjawisko małych światów



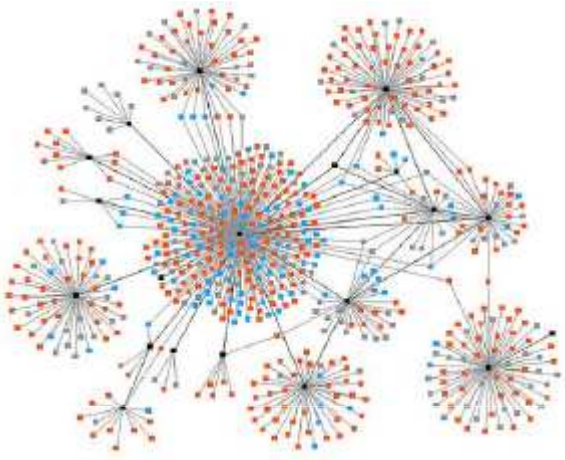
F.Karinthy, "Everything is different", 1929
M.Gurevich, M.Kochen, 1961
S.Milgram, "The Small World Problem", 1967
D.Watts, e-mail network, 2001
J.Leskovec, E. Horvitz, MS Messenger, 2006
"Six Degrees of Kevin Bacon"

Zjawisko małych światów (ang. *small world phenomenon*) występuje w sieciach, które w pierwszej chwili wyglądają na losowe, ale charakteryzują się dwoma cechami:

- posiadają stosunkowo wysoką wartość współczynnika grupowania (węzły tworzą lokalnie gęste grupy)
- posiadają niewielką średnicę i niewielką średnią odległość między węzłami (każdy węzeł może być osiągnięty w kilku krokach)

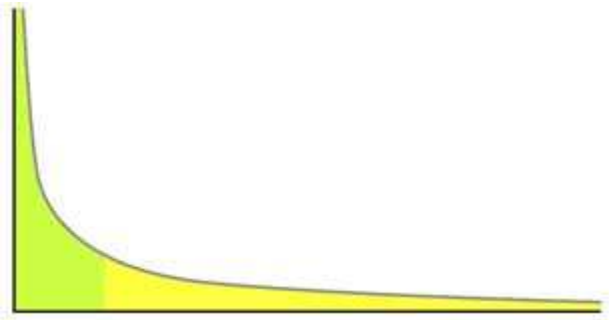
Zjawisko małych światów jest bardzo częste w sieciach społecznych ze względu na przechodniość silnych związków i tendencję słabych związków do tworzenia mostów (które skracają średnią odległość między węzłami)

Preferencyjne dołączanie



Zjawisko preferencyjnego dołączania (ang. *preferential attachement*) polega na tym, że w trakcie życia sieci nowe związki dotyczą węzłów, które posiadają wysoki stopień

- w wyniku powstaje sieć, w której niewielka liczba węzłów ma bardzo wysoki stopień a większość węzłów ma bardzo niski stopień
- sieci posiadające tę własność charakteryzują się potęgowym rozkładem stopni wierzchołków
- preferencyjne dołączanie zazwyczaj skutkuje pojawieniem się zjawiska małych światów



Powody występowania preferencyjnego dołączania mogą być różne:

- popularność
- jakość
- model mieszany

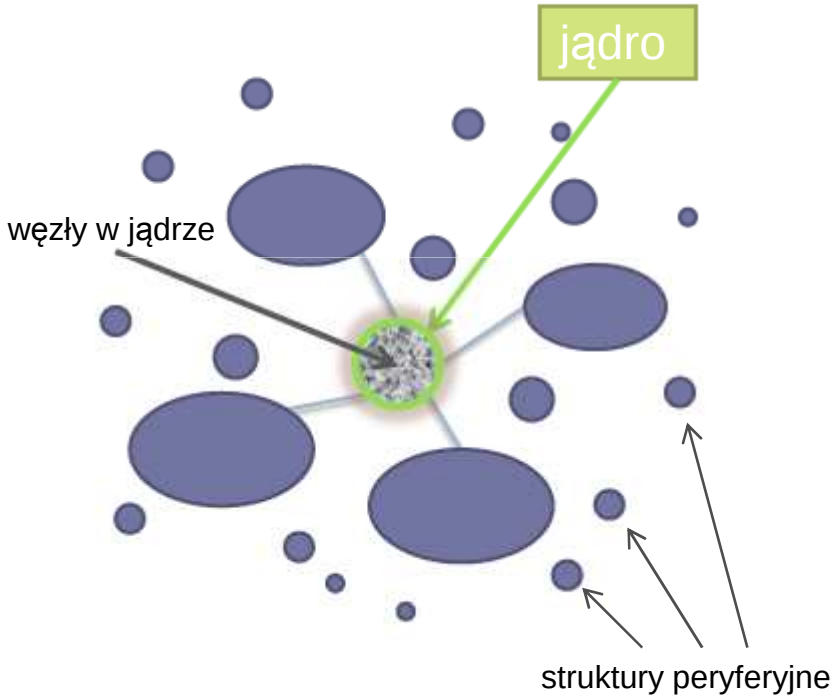
Struktury jądra i struktury peryferyjne

Wygodną metryką stopnia scentralizowania sieci jest **miara centralizacji** (ang. *centralization*)

- wyliczenie centralizacji sieci odbywa się na podstawie różnicy stopni wierzchołków
- scentralizowane sieci posiadają większość związków do wybranych węzłów, takie sieci są przydatne do rozwiązywania wielu problemów (szczególnie problemów wymagających koordynacji), ale są dużo bardziej zawodne i narażone na ataki niż sieci zdecentralizowane

Poza centralizacją wiele sieci posiada spójne silnie powiązane jądro zawierające odnośniki do struktur peryferyjnych

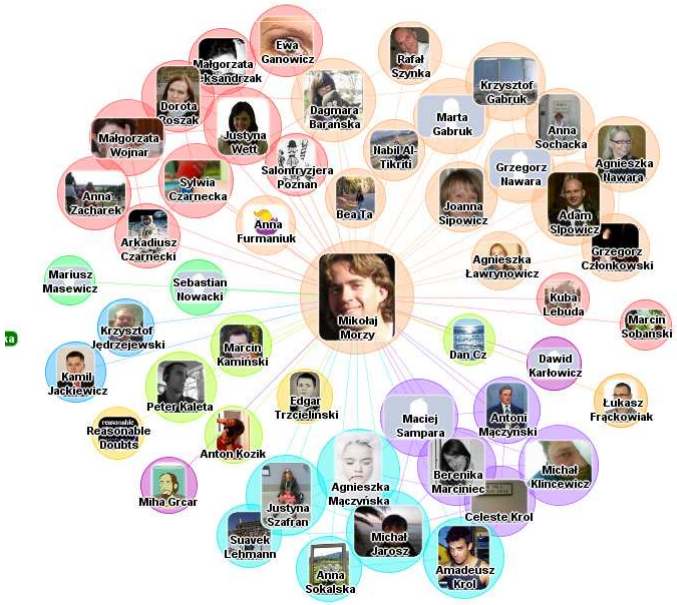
- jądro może być znalezione na podstawie analizy wizualnej czy analizy rozkładu stopni wierzchołków



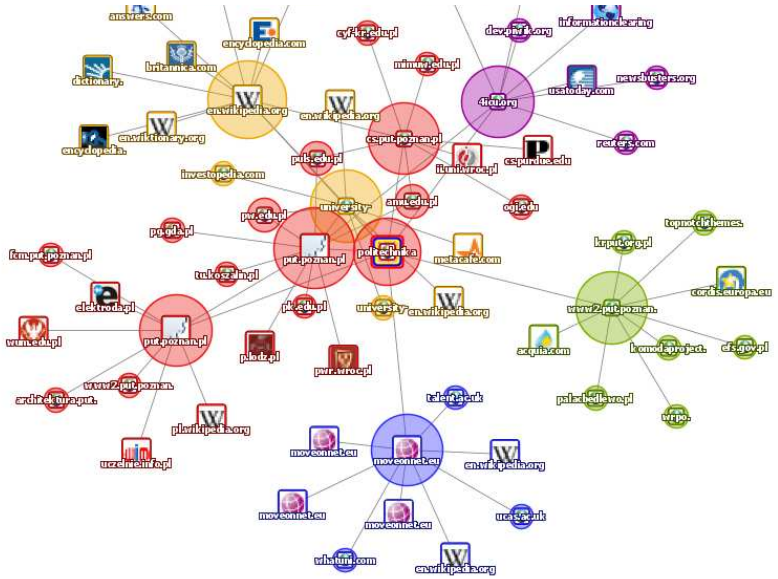
Samodzielna analiza sieci (1/3)

- uruchom TouchGraph Facebook Browser (http://apps.facebook.com/touchgraph)
- upewnij się, że zaznaczyłaś(eś) wystarczającą liczbę znajomych

- starannie zbadaj wynikową sieć
  - sprawdź rankingi znajomych oraz ich pozycję w sieci, zbadaj znalezione grupy i ich cechy charakterystyczne, postaraj się znaleźć silne i słabe związki w sieci, oceń ogólną strukturę sieci



http://www.touchgraph.com/TGGoogleBrowser.html



Samodzielna analiza sieci (2/3)

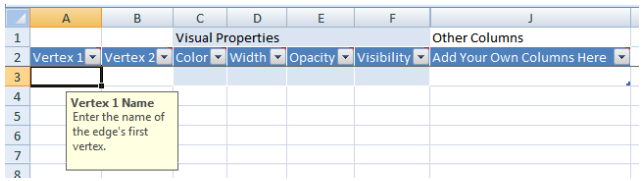
- uruchom NameGenWeb (http://apps.facebook.com/namegenweb)
- wybierz format UCInet eksportu danych
- zapisz dane na dysku

- uruchom MS Excel i otwórz wygenerowany plik
  - w automatycznie otwartym kreatorze importu plików tekstowych wskaż, że plik jest rozdzielony za pomocą znaku spacji
  - w wynikowym pliku zauważ, że najpierw znajduje się lista węzłów, a następnie lista krawędzi
  - zaznacz wszystkie krawędzie i skopiuj do schowka

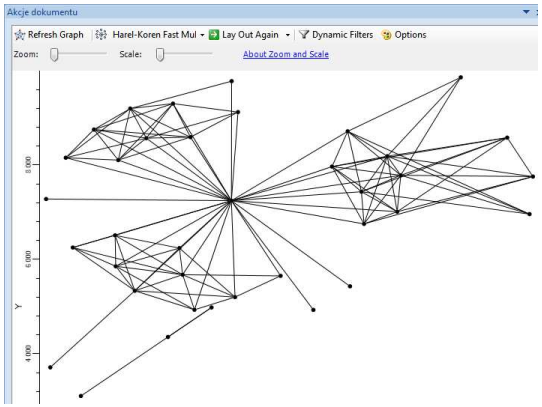
- dołącz do utworzonej na Facebooku grupy "*Technologie semantyczne i sieci społecznościowe (TSiSS)*"
  - wystarczy poszukać TSiSS
- co tydzień wykonaj zrzut aktualnej struktury Twojej sieci egocentrycznej
- na koniec semestru przygotuj opracowanie pokazujące trend zmian

Samodzielna analiza sieci (3/3)

- pobierz i zainstaluj program NodeXL (http://nodexl.codeplex.com)
- uruchom Microsoft NodeXL → Excel 2007 Template
- zaznacz pierwszą komórkę w arkuszu i skopiuj do niej listę krawędzi wygenerowaną w poprzednim kroku



- wybierz menu NodeXL → Prepare Data → Get Vertices from Edge List
- wybierz menu NodeXL → Graph Metrics i wylicz wszystkie miary
- wygeneruj graf, zbadaj jego własności, sprawdź działanie filtrów



Inne programy do analizy sieci społecznych

Na rynku jest dostępnych wiele narzędzi do analizy sieci społecznościowych. Często narzędzia te są trudne w obsłudze, mało intuicyjne, oraz wymagają wiedzy eksperckiej.

- Pajek http://vlado.fmf.uni-lj.si/pub/networks/pajek/
- UCInet http://www.analytictech.com/ucinet/
- NetDraw http://www.analytictech.com/netdraw/netdraw.htm
- GUESS http://graphexploration.cond.org/
- wiele pakietów do systemu R

- pełna lista: http://en.wikipedia.org/wiki/Social_network_analysis_software