Computational Social Science

Course #04199, module 04IN2042

How can we analyze social networks?

Markus Strohmaier

Prof. Dr.

WeST / IWVI U. of Koblenz-Landau, Germany

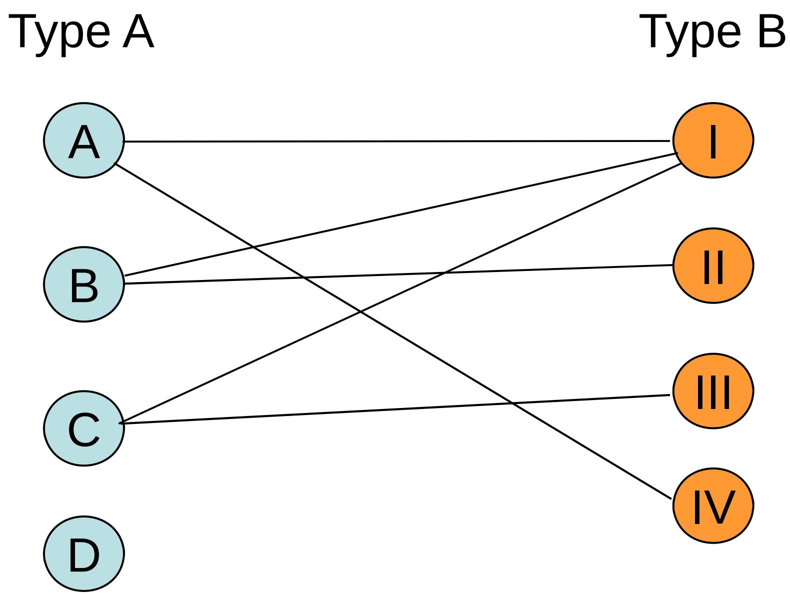
e-mail: markus.strohmaier@gesis.org
web: http://www.markusstrohmaier.info

A Small World

- Target person:
  - A Boston stockbroker
- Three starting populations
  - 100 “Nebraska stockholders”
  - 96 “Nebraska random”
  - 100 “Boston random”

Two Mode Networks

- Example:
- Two types of nodes

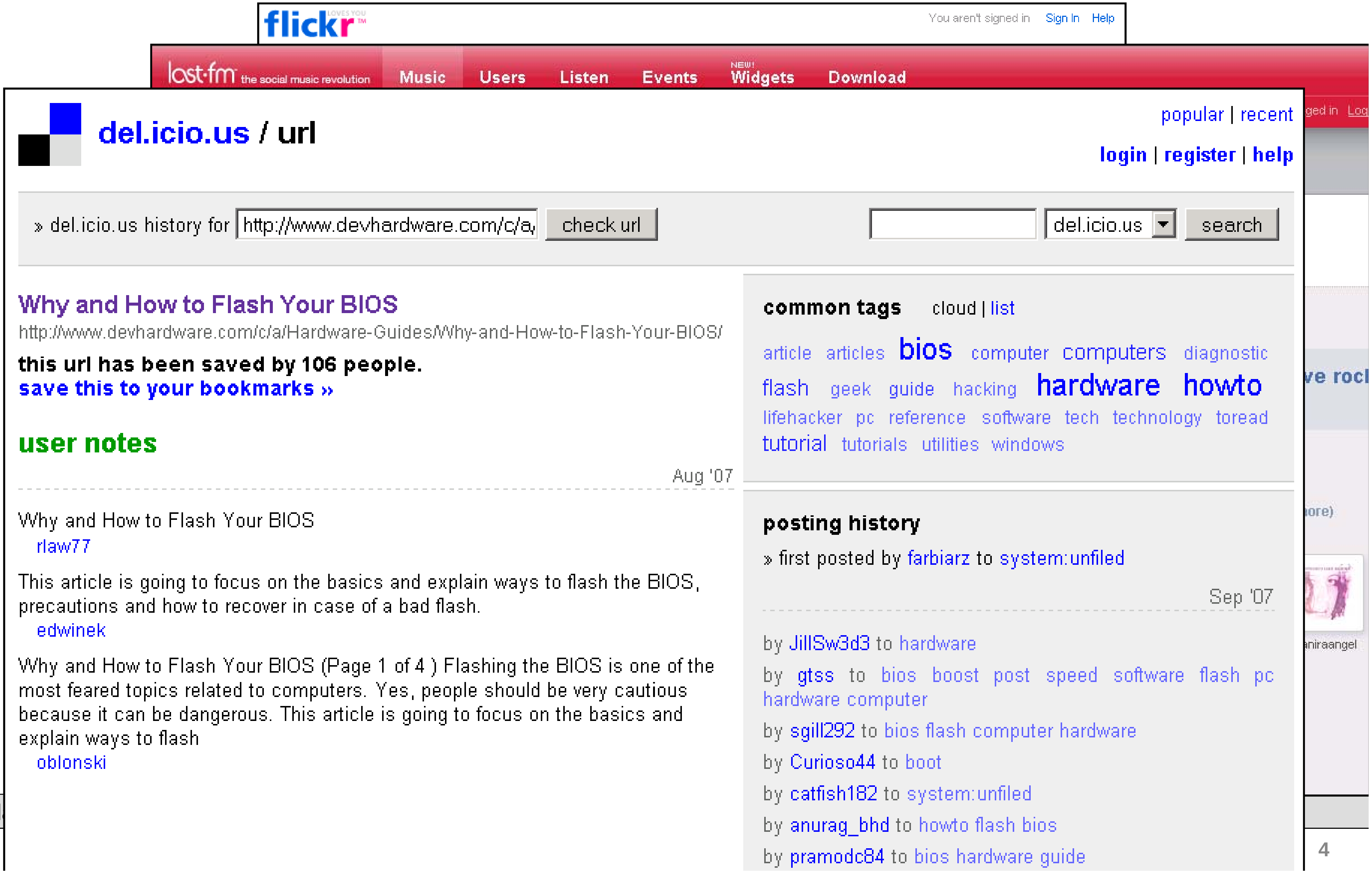


Examples: actors, scientists, students

Examples: conferences, courses, movies, articles

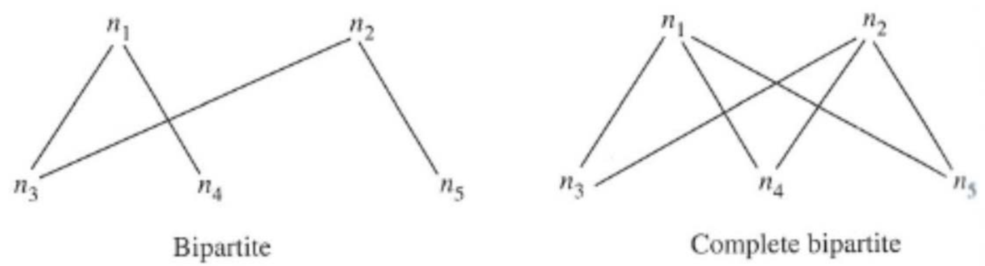
Can you give examples of two mode networks?

Reminder: Social Networks Examples



Affiliation Networks

- Affiliation networks are two-mode networks
  - Nodes of one type „affiliate“ with nodes of the other type (only!)
- Affiliation networks consist of subsets of actors, rather than simply pairs of actors
- Connections among members of one of the modes are based on linkages established through the second
- Affiliation networks allow to study the dual perspectives of the actors and the events

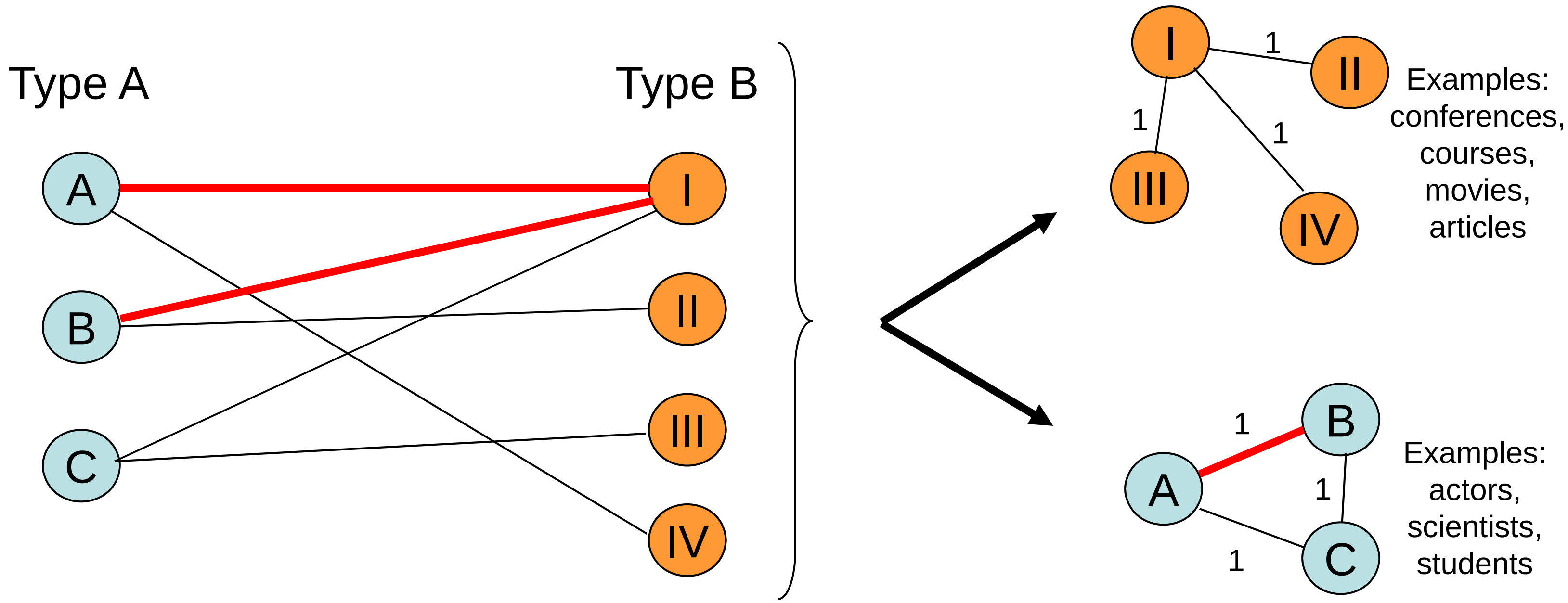


[Wasserman Faust 1994]

Fig. 4.15. Bipartite graphs

Two Mode Networks and One Mode Networks

- **Folding** is the process of transforming two mode networks into one mode networks
  - Also referred to as: **T**, ⊥ projections [Latapy et al 2006]
- Each two mode network can be folded into 2 one mode networks



Two mode network **2 One mode networks**

Today

Agenda:

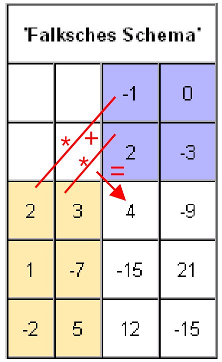
How can we analyze social networks?

A selection of concepts from Social Network Analysis

- KNC Plots
- Prominence
- Cliques, clans and clubs

Transforming Two Mode Networks into One Mode Networks

[Wasserman Faust 1994]



$$\boxed{M_P = M_{PC} * M_{PC}'}$$

C…Children

P…Party

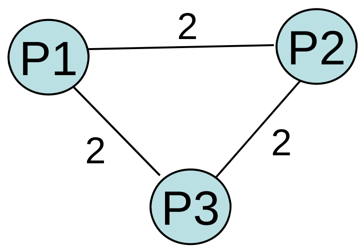
	Allison	Drew	Eliot	Keith	Ross	Sarah
Party 1	1	0	0	0	1	1
Party 2	0	1	1	0	1	1
Party 3	1	0	1	1	1	0

*

	Party 1	Party 2	Party 3
Allison	1	0	1
Drew	0	1	0
Eliot	0	1	1
Keith	0	0	1
Ross	1	1	1
Sarah	1	1	0

=

	Party 1	Party 2	Party 3
Party 1	3	2	2
Party 2	2	4	2
Party 3	2	2	4

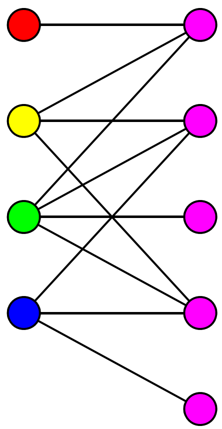


Output: Weighted regular graph

Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

The k-neighborhood graph, G_k

Given bipartite graph B, users on left, interests on right

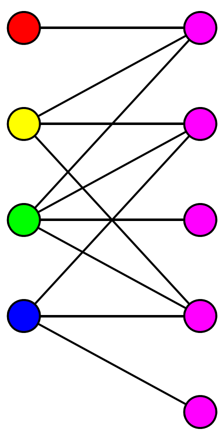


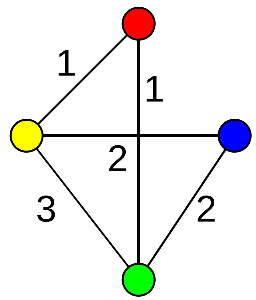
Connect two users if they share at least **k** interests in common

Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

The k-neighborhood graph, G_k

Given bipartite graph B, users on left, interests on right





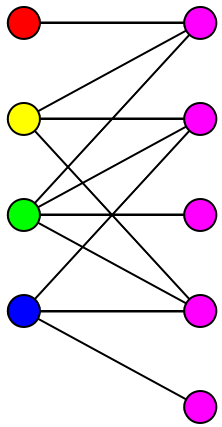
G_1

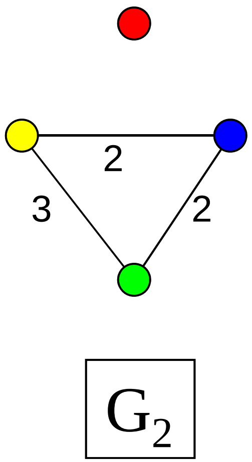
Connect two users if they share at least **k** interests in common

Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

The k-neighborhood graph, G_k

Given bipartite graph B, users on left, interests on right



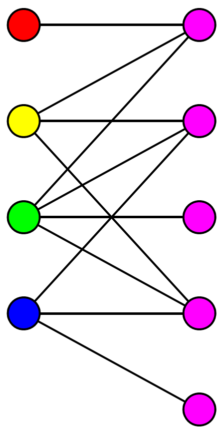


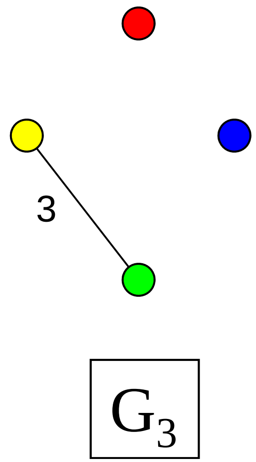
Connect two users if they share at least **k** interests in common

Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

The k-neighborhood graph, G_k

Given bipartite graph B, users on left, interests on right

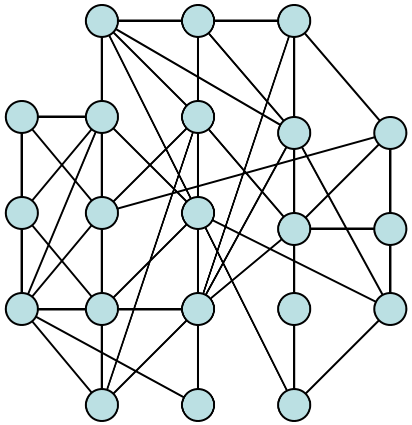




Connect two users if they share at least **k** interests in common

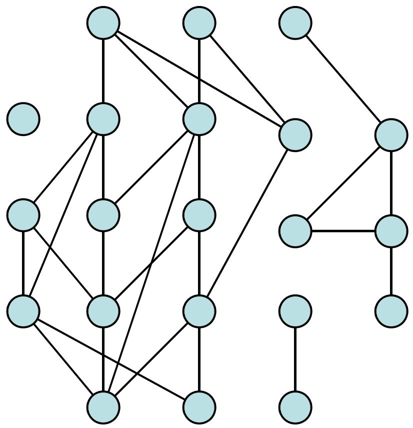
Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

Illustration k=1



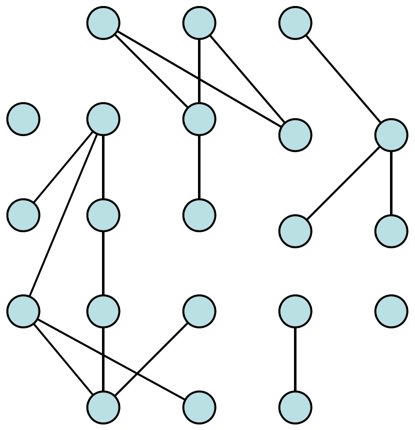
Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

Illustration k=2



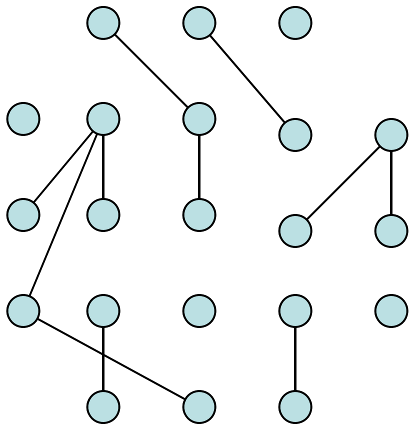
Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

Illustration k=3



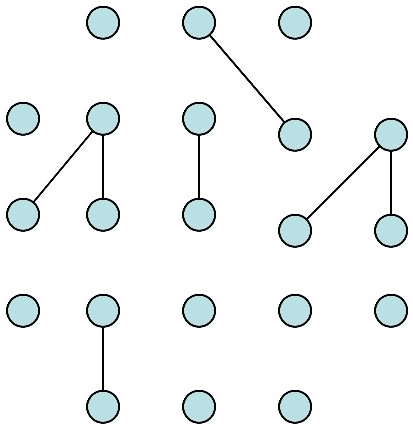
Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

Illustration k=4



Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

Illustration k=5



Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

The KNC-plot

The k-neighbor connectivity plot

- How many connected components does G_k have?
- What is the size of the largest component?

Answers the question:

how many shared interests are meaningful?

- Communities, Cuts

Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

Analysis

Four graphs:

- LiveJournal
  - Blogging site, users can specify interests
- Y! query logs (interests = queries)
  - Queries issued for Yahoo! Search (Try it at www.yahoo.com)
- Content match (users = web pages, interests = ads)
  - Ads shown on web pages
- Flickr photo tags (users = photos, interests = tags)

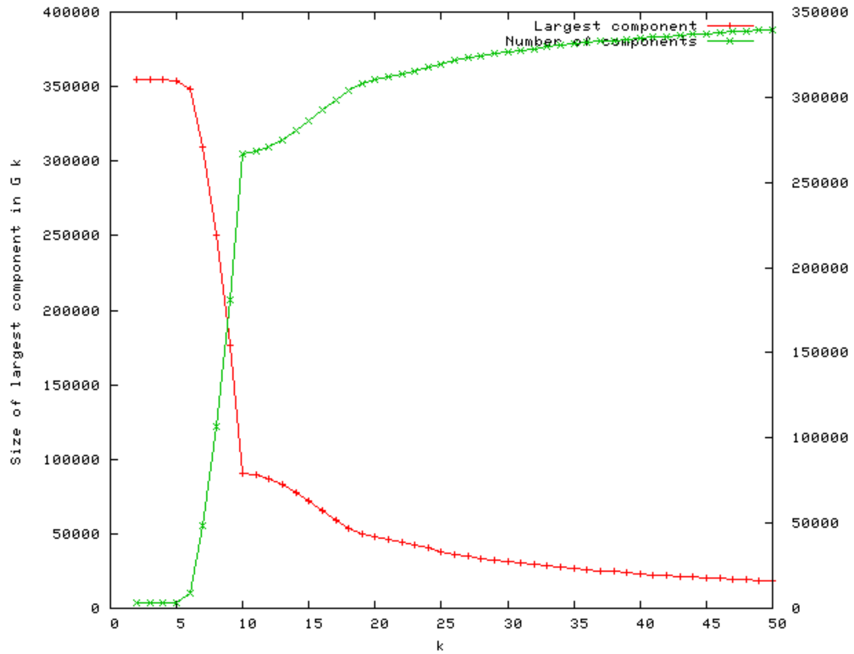
All data anonymized, sanitized, downsampled

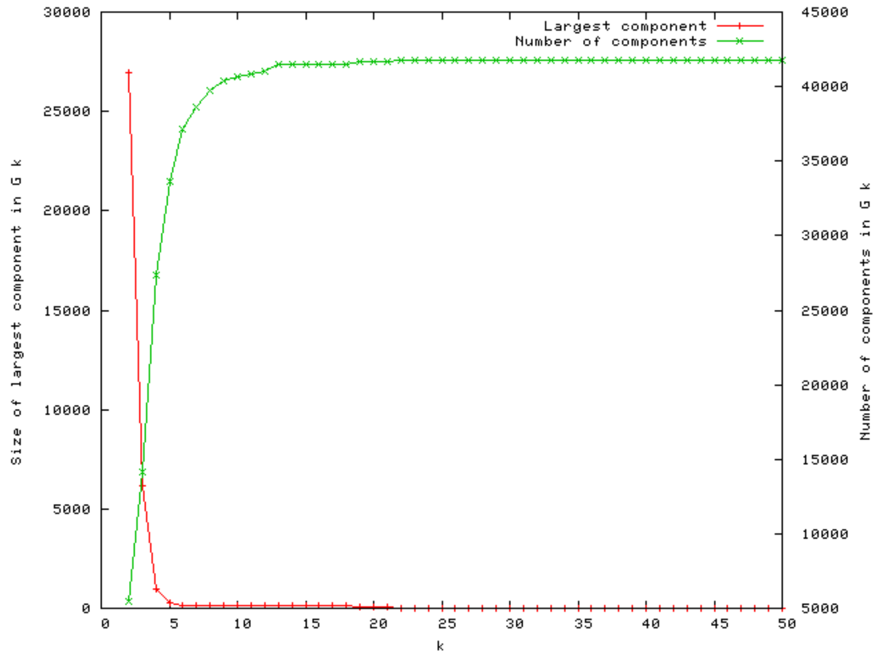
- Graphs have 100s of thousands to a million users

Slides taken from: R. Kumar and A. Tomkins and E. Vee. Connectivity structure of bipartite graphs via the KNC-plot. In Marc Najork and Andrei Z. Broder and Soumen Chakrabarti, editor(s), Proceedings of the Conference on Web Search and Data Mining, WSDM 2008, 129-138, ACM, 2008.

Examples

— Largest component
— Number of components





At k=5, all connected.
At k=6, interesting!

Content match
Web pages = “users”
Ads = “interests”

At k=6, nobody connected

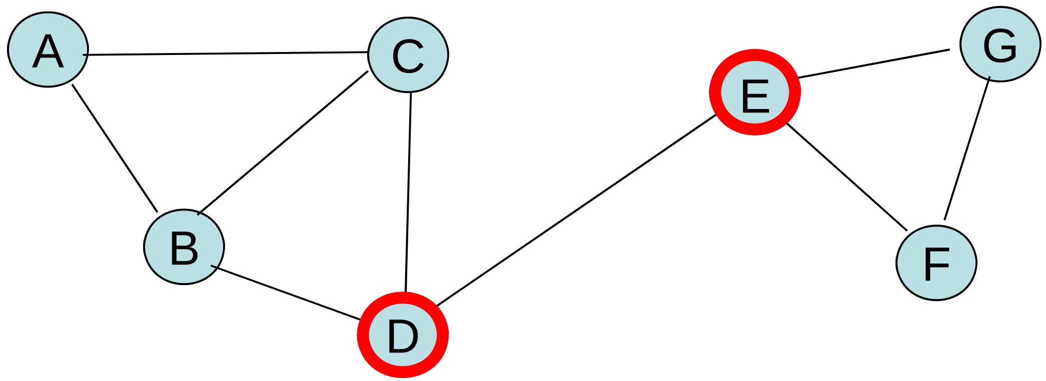
Flickr
Photos = “users”
Tags = “interests”

Cutpoint

A node, n_i, is a cutpoint if the number of components in a graph G that contains n_i is fewer than the number of components in the subgraph that results from deleting n_i from the graph.

Cutpoint or „Articulation point“

Analogous to the concept of bridges, Wasserman p113



Which node(s) represents a cutpoint? Why?

Let's take a step back!

The Web Graph is Flat

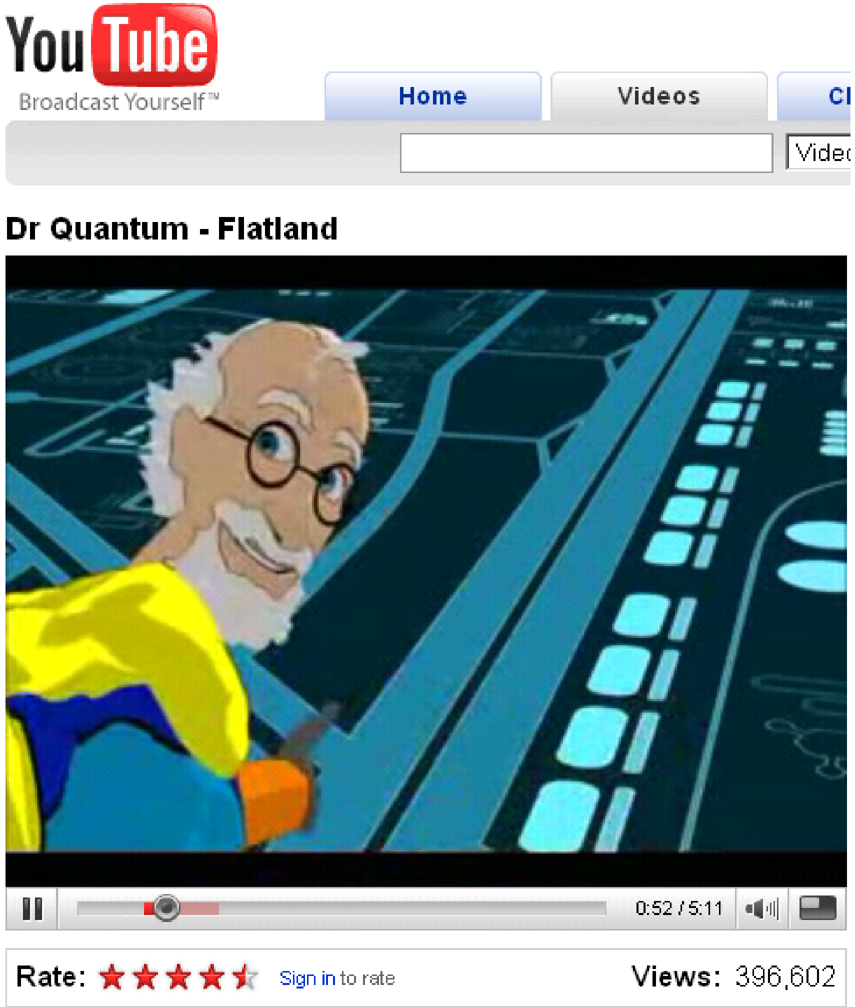
Book tip

„Flatland: A romance of many dimensions“
Edwin A. Abbott 1838-1926 (1884)

http://www.geom.uiuc.edu/~banchoff/Flatland/

How can we infer information about the n^{th}+1 dimension?

E.g. popularity, trust, prestige, importance, …

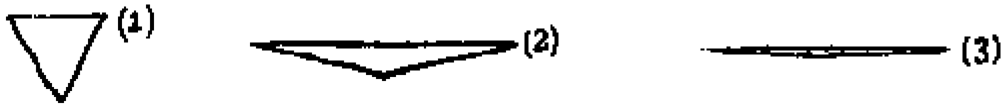


http://www.youtube.com/watch?v=BWyTxCsIXE4

http://www.flatlandthefilm.com/

Inhabitants of Flatland

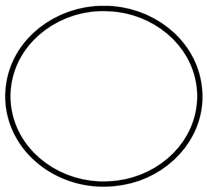
Tradesman



Men (The hero in this novel is **A. Square**)

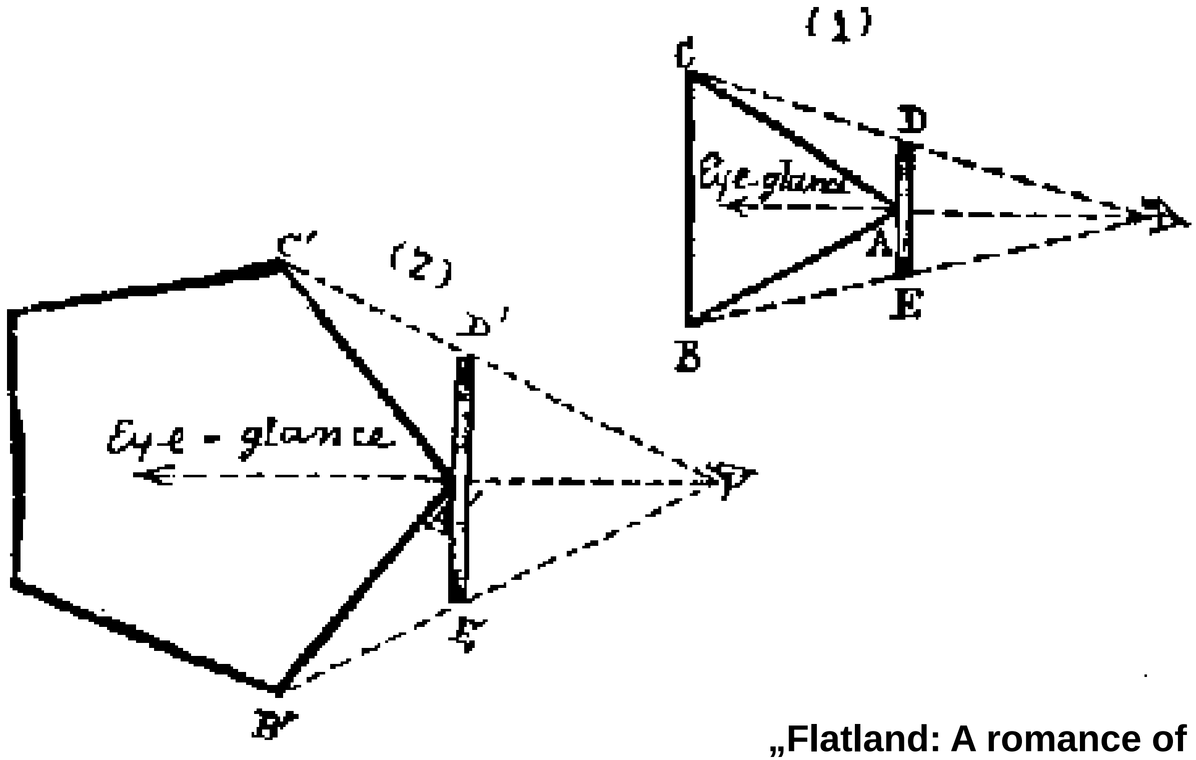
Woman

Priests



Book tip
„Flatland: A romance of many dimensions“
Edwin A. Abbott 1838-1926 (1884)
http://www.geom.uiuc.edu/~banchoff/Flatland/

Recognition by sight



Book tip
„Flatland: A romance of many dimensions“
Edwin A. Abbott 1838-1926 (1884)
http://www.geom.uiuc.edu/~banchoff/Flatland/

What kind of information can we infer from a „flat“ social graph?

Centrality and Prestige [Wasserman Faust 1994]

Which actors are the most important or the most prominent in a given social network?

What kind of measures could we use to answer this (or similar questions)?

What are the implications of directed/undirected social graphs on calculating prominence?

⇨ In directed graphs, we can use Centrality and Prestige

⇨ In undirected graphs, we can only use Centrality

Prominence
[Wasserman Faust 1994]

We will consider an actor to be prominent if the ties of the actor make the actor particularly visible to the other actors in the network.

Actor Centrality
[Wasserman Faust 1994]

Prominent actors are those that are extensively involved in relationships with other actors.

This involvement makes them more visible to the others

No focus on directionality -> what is emphasized is that the actor is involved

A *central actor* is one that is involved in many ties. [cf. Degree of nodes]

Actor Prestige
[Wasserman Faust 1994]

A prestigious actor is an actor who is the object of extensive ties, thus focusing solely on the actor as a recipient.

[cf. indegree of nodes]

Only quantifiable for directed social graphs.

Also known as *status, rank, popularity*

Different Types of Centrality in Undirected Social Graphs [Wasserman Faust 1994, Scripps et al 2007]

Degree Centrality

- Actor Degree Centrality:
  - *Based on degree only*

$$C_D(n_i) = \sum_j I[(i,j) \in E]$$

Where *I* is a 0=1 indicator function.

Closeness Centrality

- Actor Closeness Centrality:
  - Based on how close an actor is to all the other actors in the set of actors
  - Closeness is the reciprocal of the sum of all the geodesic (shortest) distances from a given node to all others
  - Nodes with a small *CC* score are closer to the center of the network while those with higher scores are closer to the edge.

$$C_C(n_i) = \left[\sum_{j=1}^{N} d(n_i, n_j)\right]^{-1}$$

d(*u; v*) is the geodesic distance from *u* to *v*.

Betweeness Centrality

- Actor Betweeness Centrality:
  - An actor is central if it lies between other actors on their geodesics
  - The central actor must be between many of the actors via their geodesics

$$C_B(n_i) = \sum_{j<k} \frac{g_{jk}(n_i)}{g_{jk}}$$

where *gjk* is the number of geodesic paths from *j* to *k (j,k all pairs of nodes)* and *gjk*(*ni*) is the number of geodesic paths from *j* to *k* that go through *i*.

→All three can be normalized to a value between 0 and 1 by dividing it with its max. value

Centrality and Prestige in Undirected Social Graphs [Wasserman Faust 1994]

Actor = closeness = betweenness centrality:

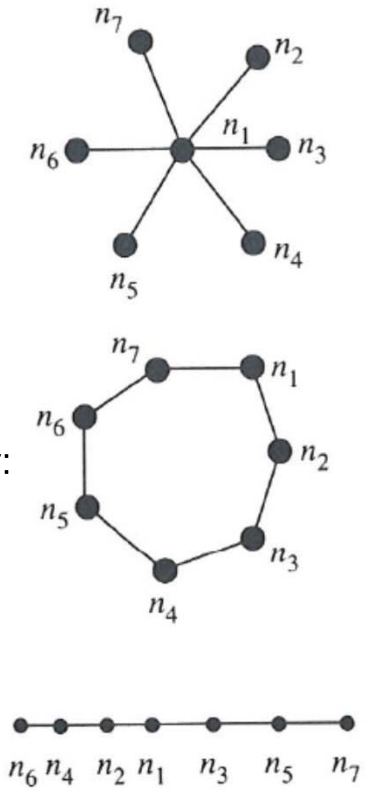
n1>n2,n3,n4,n5,n6 ,n7

Actor centrality = Betweeness centrality = Closeness centrality:

n1=n2=n3=n4=n5=n6 =n7

Betweeness centrality:

n1>n2,n3>n4,n5>n 6,n7



(a) Star graph

$$\begin{matrix} 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

(b) Circle graph

$$\begin{matrix} 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \end{matrix}$$

(c) Line graph

$$\begin{matrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{matrix}$$

Fig. 5.1. Three illustrative networks for the study of centrality and prestige

Examples and Simulation:

How can we identify groups and subgroups in a social graph?

Cliques, Subgroups [Wasserman Faust 1994]

How can we identify groups and subgroups in a social graph?

What cliques can you identify in the following graph?

Definition of a Clique

- A clique in a graph is a maximal *complete* subgraph of three or more nodes.

Remark:

- Restriction to at least three nodes ensures that dyads are not considered to be cliques
- Definition allows cliques to overlap

Informally:

- A collection of actors in which each actor is adjacent to the other members of the clique

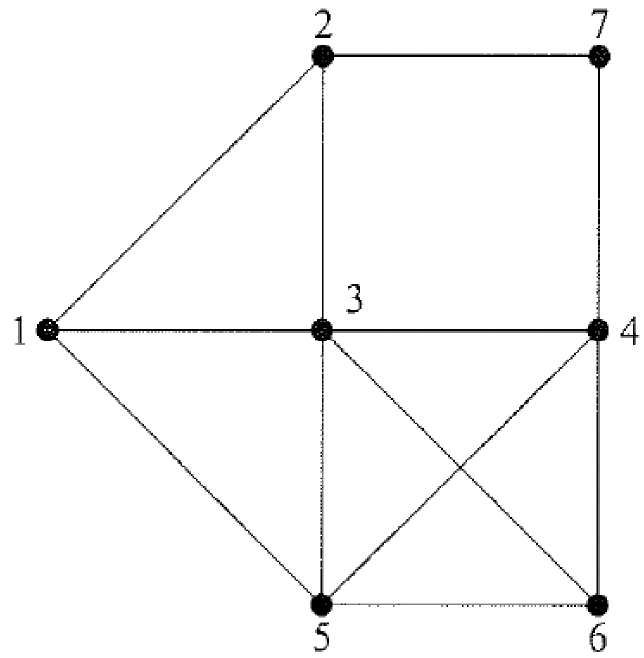


Fig. 7.1. A graph and its cliques

Subgroups
[Wasserman Faust 1994]

Cliques are very strict measures

- Absence of a single tie results in the subgroup not being a clique
- Within a clique, all actors are theoretically identical (no internal differentiation)
- Cliques are seldom useful in the analysis of actual social network data because definition is overly strict

⇨ So how can the notion of cliques be extended to make the resulting subgroups more substantively and theoretically interesting?

⇨ Subgroups based on reachability and diameter

n cliques
[Wasserman Faust 1994]

N-cliques require that the **geodesic distances** among members of a subgroup **are small** by defining a **cutoff value n** as the maximum length of geodesics connecting pairs of actors within the cohesive subgroup.

An n-clique is a maximal ~~*complete*~~ subgraph in which the largest geodesic distance between any two nodes is no greater than n.

Which 2-cliques can you identify in the following graph?

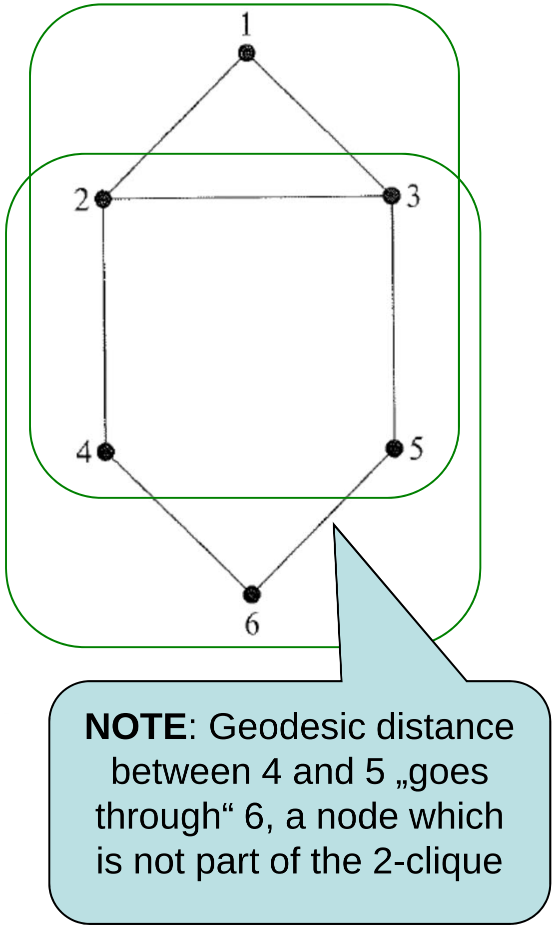


Fig. 7.2. Graph illustrating *n*-cliques, *n*-clans, and *n*-clubs

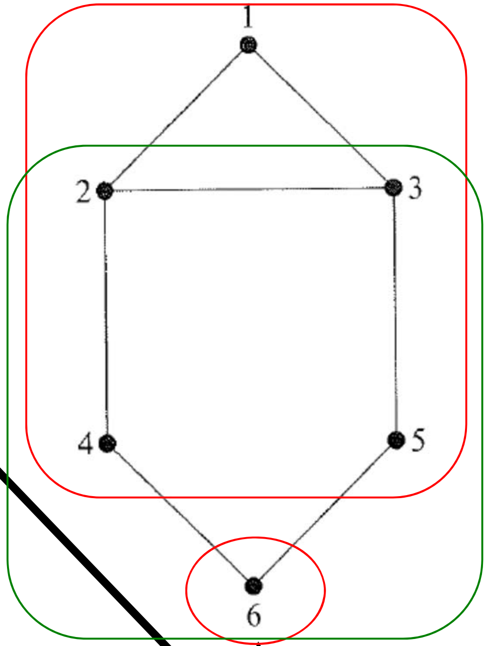
n clans
[Wasserman Faust 1994]

An n-clan **is an n-clique** in which the geodesic distance between all nodes in the subgraph is no greater than n for paths **within** the subgraph.

N-clans in a graph are **those n-cliques** that have diameter less than or equal to n (within the graph).

⇨ All n-clans **are** n-cliques.

Which 2-clans can you identify in the following graph?



Why is {1,2,3,4,5} not a 2-clan?

Why is {1,2,3,4} not a 2-clan?

Fig. 7.2. Graph illustrating n-cliques, n-clans, and n-clubs

n clubs
[Wasserman Faust 1994]

An n-club is defined as a maximal subgraph of diameter n.

No node can be added without increasing the diameter.

A subgraph in which the distance between all nodes **within the subgraph** is less than or equal to n

And no nodes can be added that also have geodesic distance n or less from all members of the subgraph

⇨ All n-clubs are **contained within** n-cliques.

⇨ All n-clans are also n-clubs

⇨ Not all n-clubs are n-clans

Which 2-clubs can you identify in the following graph?

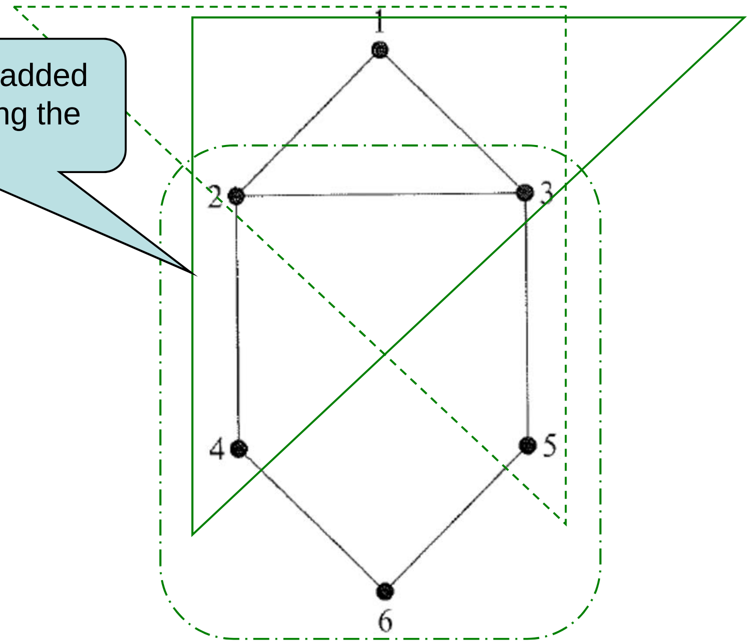


Fig. 7.2. Graph illustrating n-cliques, n-clans, and n-clubs

K-cores

Include a discussion of k-cores, as a way to reduce „noise links" in a graph, while at the same time keeping important subgroups/clusters in a **weighted** graph

Used by Barry Wellman / Drew Conway (there's a post on the zero intelligence blog)

Implemented by Network X

TODO

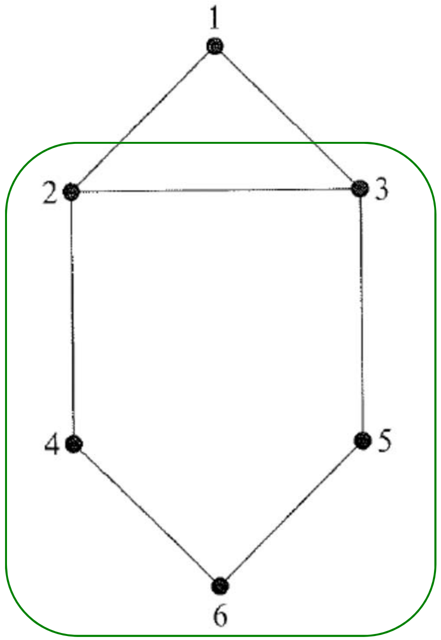
K-plexes [Wasserman Faust 1994]

A k-plex is a maximal subgraph containing N nodes in which each node is adjacent to no fewer than N-k nodes in the subgraph.

In other words, each node in the subgraph may be lacking ties to no more than k subgraph members.

A clique occurs when k=1.

Can you identify a 3-plex in the following graph?



Subgroups in Co-Affiliation Networks
Borgatti 1997

- The obvious next step would be to try to identify these subgroups in co-affiliation networks.
  - For example, we can search for cliques, n-cliques, n-clans, n-clubs.
- Unfortunately, these methods are not well suited for analysing a bipartite graph.
  - In fact, bipartite graphs contain no cliques
  - In contrast, bipartite graphs contain too many 2-cliques and 2-clans.
  - One of the problems is that, in the bipartite graph, all nodes of the same type are necessarily two links distant.

➔ we need to consider special types of subgraphs which are more appropriate for two-mode data.

Bicliques
[Borgatti 1997]

- A biclique is a maximal complete bipartite subgraph of a given bipartite graph.
- Reasonable to insist on bicliques of the form $K_{m,n}$ where m and n are greater than 2
  - Why? Each of the two modes should form (after folding) interesting structures (triads or greater)

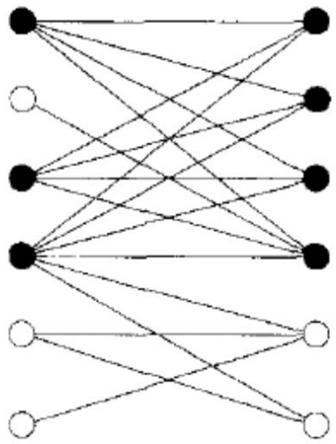
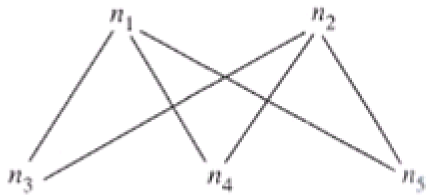


Fig. 10. Dark nodes form a biclique.



Complete bipartite

Wasserman / Faust 1994

Subgroups in Co-Affiliation Networks
Borgatti 1997

- Clearly, we can define extensions of n-cliques, n-clubs and n-clans to n-bicliques, n-biclubs and n-biclans.
- But, the extensions would in many senses be unnatural since n would need to be odd.

- **There is a way to analyze subgroups in affiliation networks: Galois Lattices**

Home Assignment

- How many have started working on their assignments yet?

- In case of any questions, do not hesitate to post to the newsgroup infko.compsocsci

Any questions?

See you <u>next week</u>!