Social Network Analysis

Course week 2 – Basics of SNA, descriptives, visualization

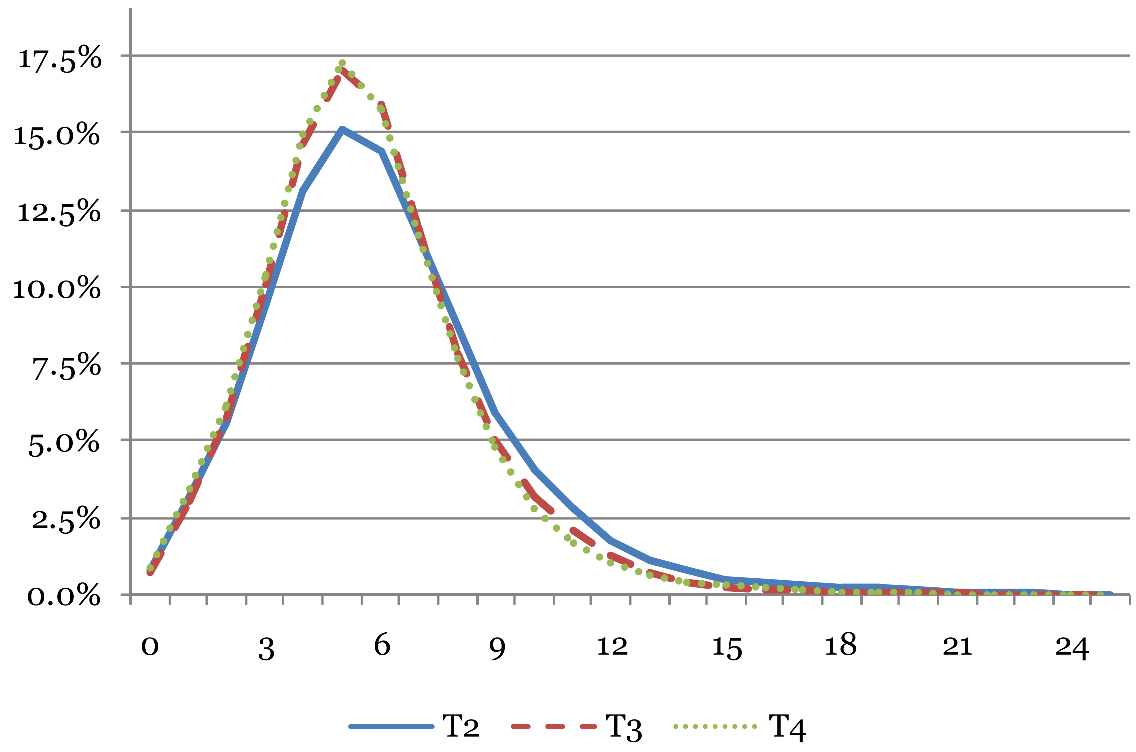
karoly.takacs@liu.se

LINKÖPING UNIVERSITY

Network level: geodesic distributions

For each network, you can calculate how often each distance occurs.

- This generates the distribution of geodesic distances.
- Actors in separate components have infinite distance.



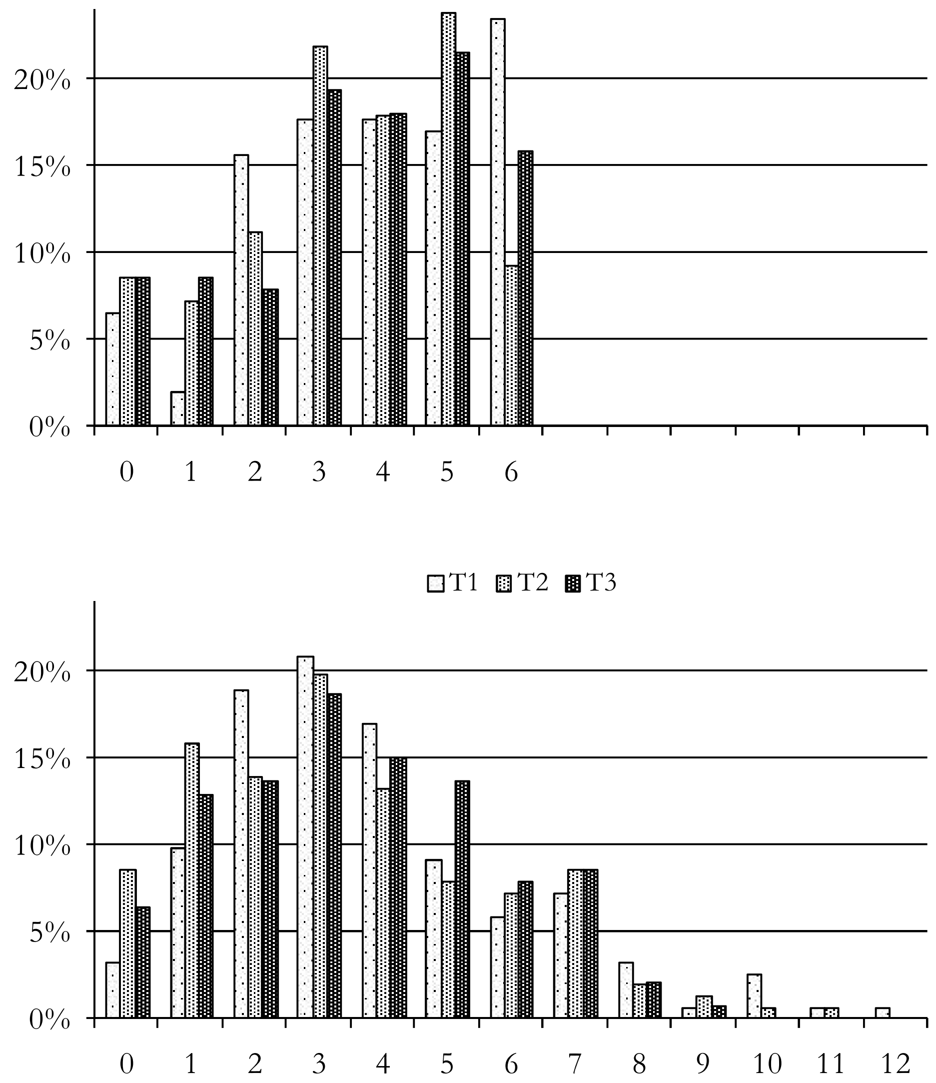
On the right you see a series
of distributions of geodesic distances
in English/Welsh school-based friendship networks over 3 years.

.. & degree distributions

For each network, you can calculate how often each out- or in-degree occurs.

- This generates the so-called degree distributions.
- Outliers on the right are called „hubs“.

On the right you see a series of distributions of outdegrees (top) and indegrees (bottom) in a Scottish school-based friendship network over 3 years.

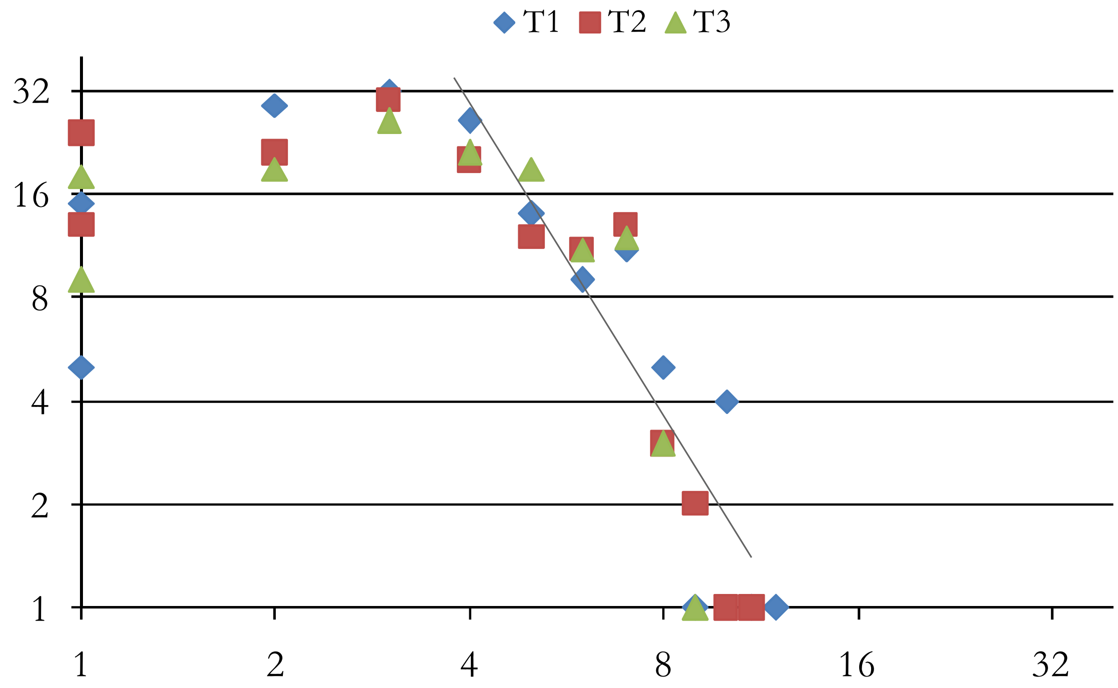


Scale-free networks

Degree distributions are often depicted on a 'log-log-scale' (physics tradition)

- If the distribution is linear on this scale, the network is called 'scale-free' (A.-L. Barabási)

On the right, you see the indegree distribution of the previous slide's data set on a log-log-scale.



Actor-level centrality notions

Basic idea behind these concepts:

Well-connected actors are in a structurally advantageous position.

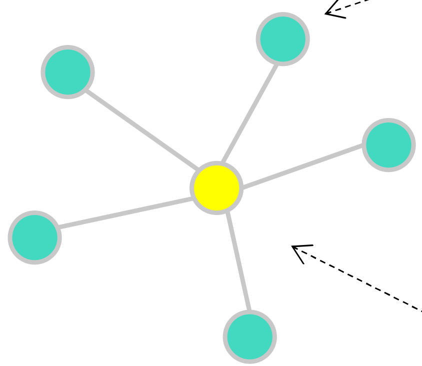
- What makes an actor “well-connected”?

  $\rightarrow$ different notions of ***centrality*** try to capture such aspects

- When actors in a network differ strongly in terms of their centrality, the network is called ***centralized*** (have high centralization)

Structurally advantageous positions

- A “star structure” expresses the centrality concept most clearly:



Peripheral actors are only indirectly connected to each other, they “rely on” the central actor for access to each other.

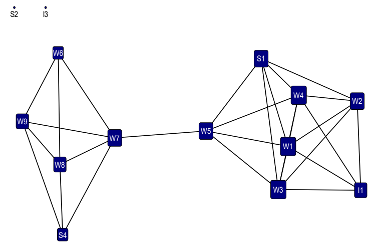
Central actor is connected to everyone, dominates access.

- There is some nuance to this…

Three actor-level centrality measures

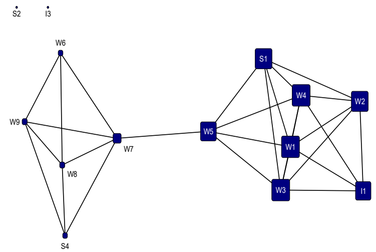
Degree centrality

- Number of people nominating you (indegree centrality) or nominated by you (outdegree centrality)



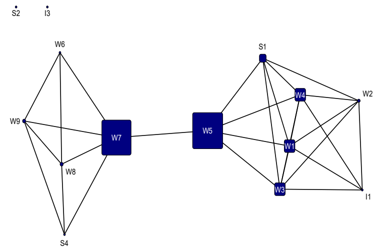
Eigenvector centrality

- Your centrality is proportional to the sum of your nominees' / nominators' centrality.

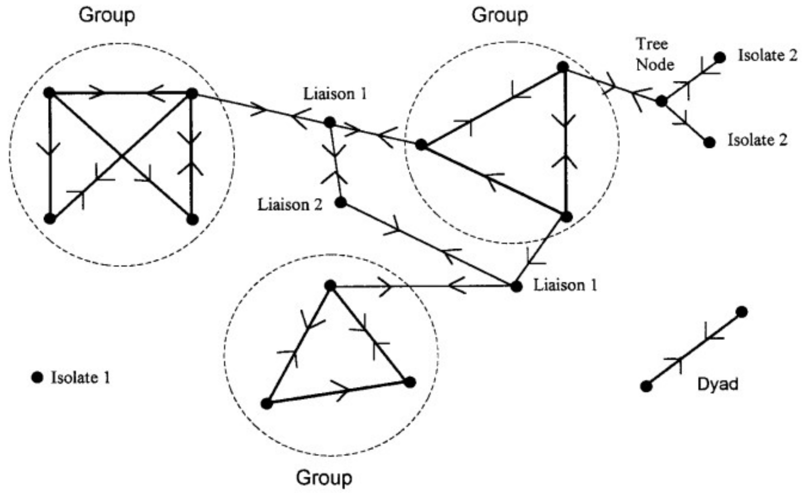


Betweenness centrality

- Sum of fractions of shortest paths between any two nodes that pass through a given node



Actor level positional measures

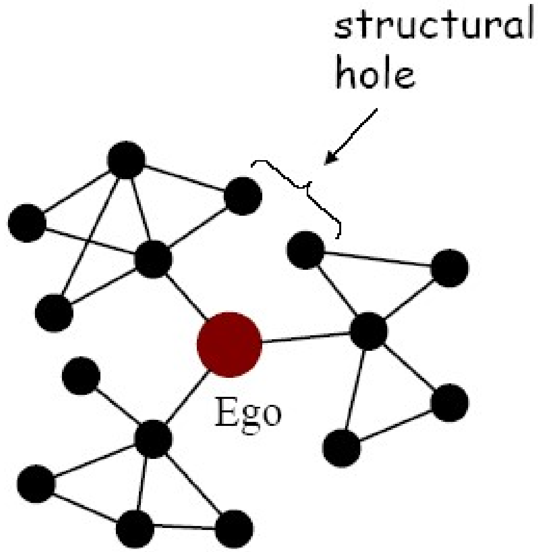


Group member – peripheral – bridge – isolate

- Problem: various definitions

Structural holes / brokerage

- Actors can exploit other actors' disconnectedness



Equivalence (later in the course)

Further network-level measures

Segmentation

- *Degree to which network falls apart into components.*

Segregation

- *Degree to which network falls apart into components that are homogeneous on some actor property (e.g., sex, race).*

Hierarchization, diameter, …

Handling network data

- Getting used to handling (and thinking about) network data: relations!
- From real-world observations to network data
- Learning standard transformations for relational data
- Insight into more advanced network data (and handling/transformations)
- 2-mode/bipartite networks

Transformations

- Types of relational data
- Network metrics & methods (mostly binary)
- Standard transformations (Dichotomization, symmetrization)
- More transformations (normalizations)
- Advanced network data structures (multiplex, multilevel)
- 2-mode networks (projections, N-mode)

Recapping the fundamentals

- Actor (node/vertex/entity etc)
  - A person, organization, country, concept, journal, group, event etc
  - Attribute: property of actor (e.g., age, name, gender, etc.)
- Relation (edge/tie/arc/connection)
  - An association/phenomena that tie two actors together (e.g., friendship, war, trade, collaboration, attendance, pipeline, road, kinship, etc.)

Network methods and metrics

- Predominantly intended for binary networks
  - i.e., where the ties either exist (1) or not (0)
- Far less methods & metrics intended for valued networks
  - Multiple dilemmas with valued networks
  - This course: primarily binary networks
- Many real-world observations are "more than binary"

Types of relations

Dichotomization

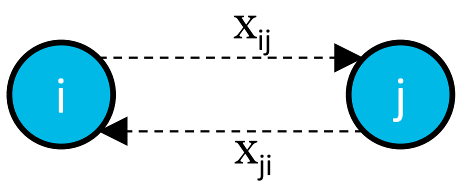
Symmetrization

	Binary	Valued
Symmetrical	• John and Sue know each other • USA and Canada have a treaty • Orgs A and B share resources	• John and Sue have met 4 times • Travel distance Malmö-Lund is 20 km • Party A and B have co-sponsored 8 bills 4
Directional	• John likes Sue • Sue dislikes John • Sweden invaded Denmark	• John gave Sue 5 apples • 3 politicians left party A for party B • John called Sue twice 5

(Also *signed* networks, where ties are + , absent, or -)

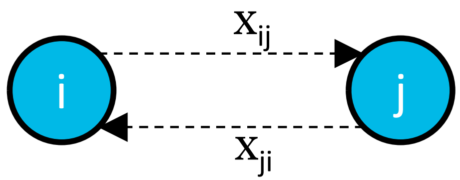
Symmetrization

- Transforming binary directional network $\boldsymbol{X}$ to symmetrical version $\boldsymbol{S}$



- Reciprocal ties: x_{ij} vs. x_{ji}
- Two possibilities:
  - Min-symmetrization (AND-logic): $s_{ij} = s_{ji} = \min(x_{ij}, x_{ji})$
    - Min-symm (AND-logic): at least an A-tie (asymmetrical)
  - Max-symmetrization (OR-logic): $s_{ij} = s_{ji} = \max(x_{ij}, x_{ji})$
    - Max-symm (OR-logic): must be an M-tie (mutual)
- (For valued networks: also sum- and mean-based symmetrizations – what makes sense depends on context)

Symmetrization



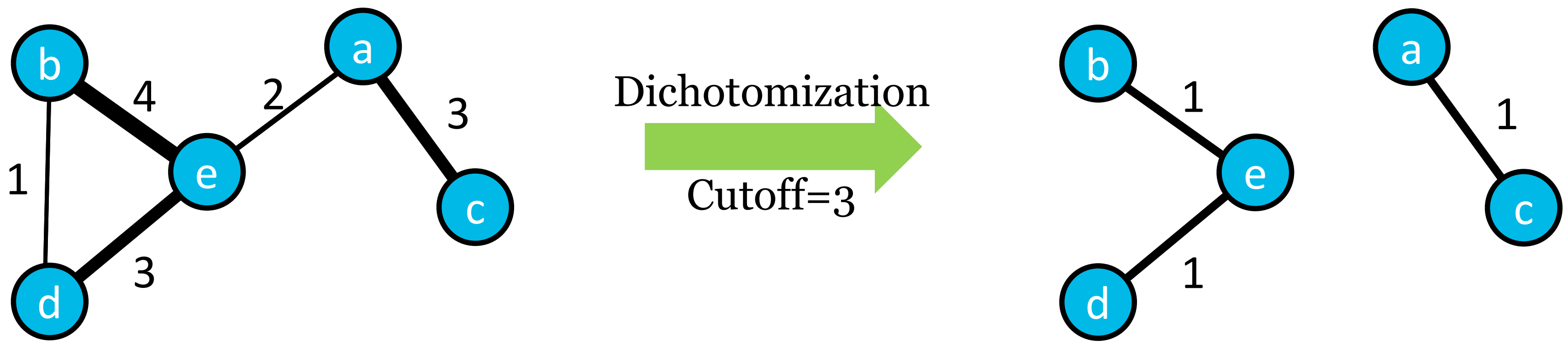
- Does symmetrization always make sense?
  - Depends on data, context, RQs!
- Reported friendships:
  - Symmetrization could be meaningful
- Advice-seeking networks:
  - Unlikely to be meaningful (typically anti-symmetrical)
- Number of sent emails (valued)
  - Total number of emails exchanged

    (sum-symmetrization)

Dichotomization

- Transforming valued network $\boldsymbol{X}$ to binary network $\boldsymbol{B}$
- Choosing *cutoff* value: defines prominent ties:

$$x_{ij} \geq cutoff \rightarrow b_{ij} = 1$$
$$x_{ij} < cutoff \rightarrow b_{ij} = 0$$



Dichotomization

- How to choose a suitable *cutoff*?
  - Statistically (e.g., top p% of total values)
  - Theoretically (e.g., inter-personal interaction exceeding 5 minutes are prominent)
  - Substantially relevant (e.g., Likert-scale on friendships: we only want “good friends” or more, not “acquaintances”)
  - All! Incremental cutoff, generate multiple dichotomized networks, compare results

Sometimes “trichotomization” makes sense, too

- e.g., *absent* , *weak,* and *strong* ties (Granovetter)

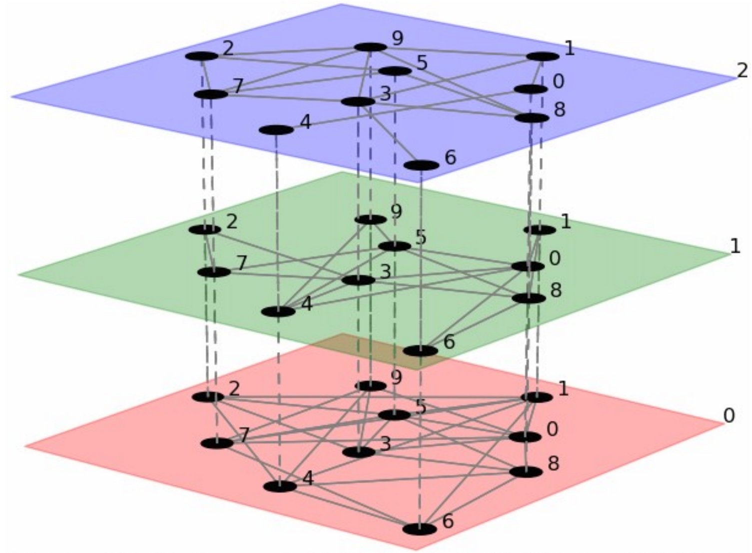
Dichotomization

- Unequal relational capacities: dichotomization often meaningless
- Schoolkids interacting during a lunch break (symmetrical; value=time interacting)
  - Each kid has 45 minutes to distribute among peers
  - System-wide cutoff thus reasonable: similar perception of prominent tie
- International trade
  - Vastly different relational capabilities
  - E.g., trade from Colombia to USA: 43% of Colombia's total exports, only 0.9% of US total imports
  - Perceptions of what is a prominent tie differs enormously among actors
- Dichotomization: strength-of-ties or patterns-of-ties?

Normalization

- Either for individual actors and their ego networks (i.e., rows and columns in the sociomatrix)…
- ….or for complete networks (which corresponds to multiplying all tie values with a scalar)
- Different types: marginal (or sum), mean etc. Marginal most common:
  - Each value is rescaled such that the sum equals unity
  - For ego-network: normalize rows (or columns) by dividing each cell x_{ij} with row (column) sum
- Examples of applicable situations:
  - Gift giving (unequal gift giving capabilities)
  - Time spent socializing (different social capital)
  - Rows or columns? Depends on context and RQ

Multiplex networks



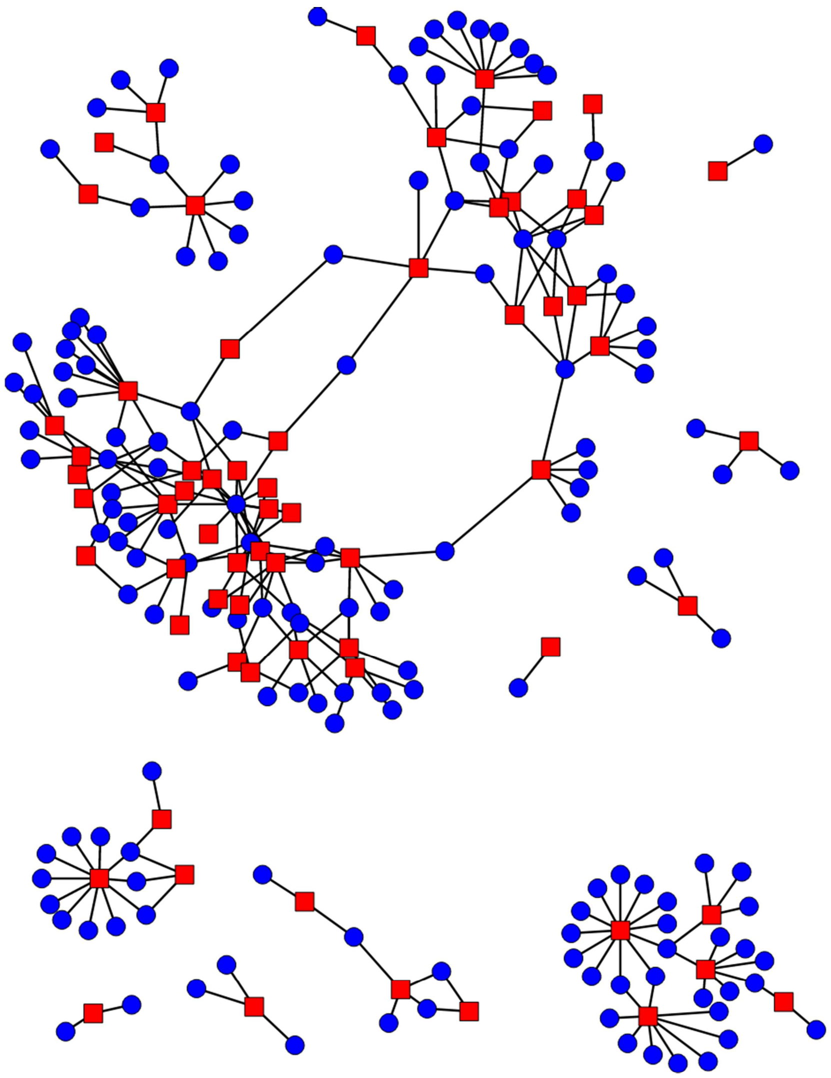
- One set of actors
  - e.g., company employees
- Multiple *types* of relations
  - Friendship
  - Advice-seeking
  - Aversion
- Many methods/metrics have versions for multiplex networks
- Comparing network types for same set of actors (e.g., QAP)

2-mode (bipartite) networks

- “Conventional” networks (“unipartite”, “1-mode”):
  - Single set of actors/nodes
  - Relations between these actors
- Affiliation networks (“bipartite”, “2-mode”):
  - Two different sets of actors/nodes
  - Relations only between nodes from different sets!
- Examples:
  - People on corporate boards
  - Co-authored articles
  - Legal references in documents
  - Political parties in parliamentary committees
  - Texas society ladies at social events

2-mode network data

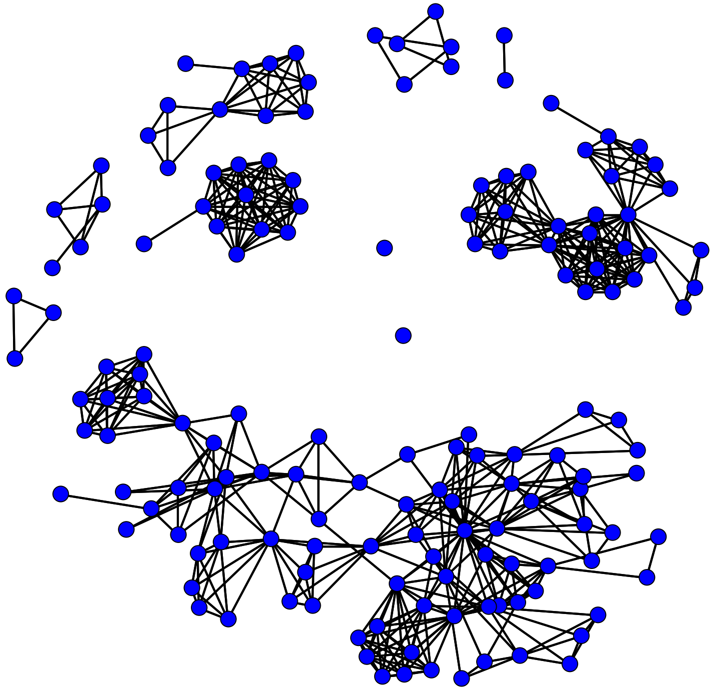
- Two-mode (a.k.a. *bipartite*) networks consist of ties between two qualitatively distinct sets of nodes, often *events* and *actors* attending them
- They can be projected to two one-mode networks: between actors (relation: sharing events) and between events (relation: sharing actors)
- *On the right, you see 'hangjongeren' (adolescents, blue circles) by police-registered incidents (red squares) in a village in province Drenthe*



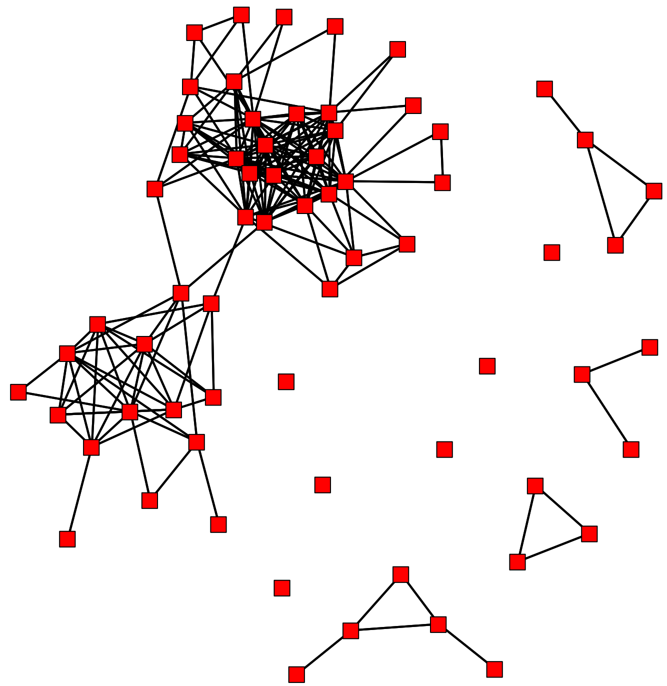
Projections of 2-mode data to one-mode data

Projections are valued networks of co-occurrence counts

How many incidents do two adolescents share ?



They can reveal a sort of 'community structure' for both types of nodes.

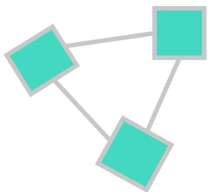


How many adolescents do two incidents share?

Non-uniqueness of projections

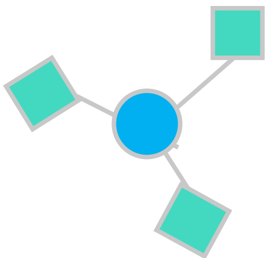
Non-uniqueness of origin (in particular after binarising)

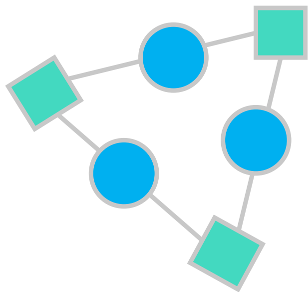
Example: the projection



could be obtained either

from this 2-mode network, or from this one:





When observing a one-mode clique structure in a projected network, this can come from a two-mode star structure (left) or from three independent links (right).

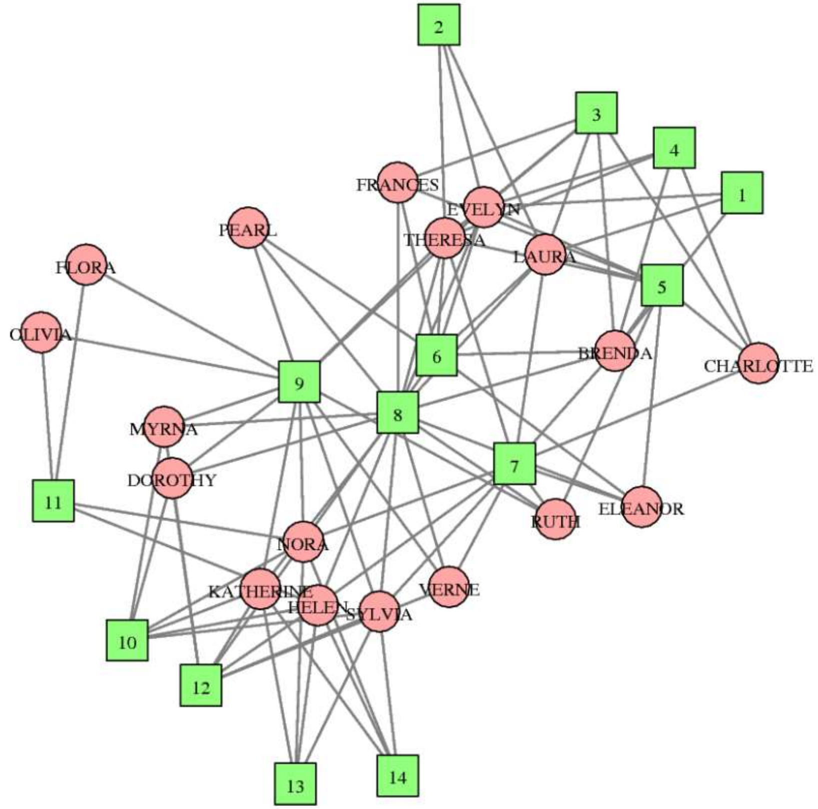
A lot of research on co-authorship networks ignores the first possibility and (wrongly) interprets the upper triangle as transitivity…

Texas society ladies

18 ladies attending 14 events

X:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ID	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
2	EVELYN	1	1	1	1	1	1	0	1	1	0	0	0	0	0
3	LAURA	1	1	1	0	1	1	1	1	0	0	0	0	0	0
4	THERESA	0	1	1	1	1	1	1	1	1	0	0	0	0	0
5	BRENDA	1	0	1	1	1	1	1	1	0	0	0	0	0	0
6	CHARLOTTE	0	0	1	1	1	0	1	0	0	0	0	0	0	0
7	FRANCES	0	0	1	0	1	1	0	1	0	0	0	0	0	0
8	ELEANOR	0	0	0	0	1	1	1	1	0	0	0	0	0	0
9	PEARL	0	0	0	0	0	1	0	1	1	0	0	0	0	0
10	RUTH	0	0	0	0	1	0	1	1	1	0	0	0	0	0
11	VERNE	0	0	0	0	0	0	1	1	1	0	0	1	0	0
12	MYRNA	0	0	0	0	0	0	0	1	1	1	0	1	0	0
13	KATHERINE	0	0	0	0	0	0	0	1	1	1	0	1	1	1
14	SYLVIA	0	0	0	0	0	0	1	1	1	1	0	1	1	1
15	NORA	0	0	0	0	0	1	1	0	1	1	1	1	1	1
16	HELEN	0	0	0	0	0	0	1	1	0	1	1	1	0	0
17	DOROTHY	0	0	0	0	0	0	0	1	1	0	0	0	0	0
18	OLIVIA	0	0	0	0	0	0	0	0	1	0	1	0	0	0
19	FLORA	0	0	0	0	0	0	0	0	1	0	1	0	0	0



(Davis, A. et al, 1941)

Note: Ties (binary, symmetric) only between nodes of *different* types

Classical 2-mode data set: Davis’ “Southern Women”

		E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14
1	EVELYN	1	1	1	1	1	1	0	1	1	0	0	0	0	0
2	LAURA	1	1												
3	THERESA	0	1												
4	BRENDA	1	0												
5	CHARLOTTE	0	0	1	1	1	0	1	0	0	0	0	0	0	0
6	FRANCES	0	0	1	0	1	1	0	1	0	0	0	0	0	0
7	ELEANOR	0	0	0	0	1	1	1	1	0	0	0	0	0	0
8	PEARL	0							1	1	0	0	0	0	0
9	RUTH	0							1	1	0	0	0	0	0
10	VERNE	0							1	1	0	0	1	0	0
11	MYRNA	0	0	0	0	0	0	0	1	1	1	0	1	0	0
12	KATHERINE	0	0	0	0	0	0	0	1	1	1	0	1	1	1
13	SYLVIA	0	0	0	0	0	0	1	1	1	1	0	1	1	1
14	NORA	0	0	0	0	0	1	1	0	1	1	1	1	1	1
15	HELEN	0	0	0	0	0	0	1	1	0	1	1	1	0	0
16	DOROTHY	0	0	0	0	0	0	0	1	1	0	0	0	0	0
17	OLIVIA	0	0	0											
18	FLORA	0	0	0											

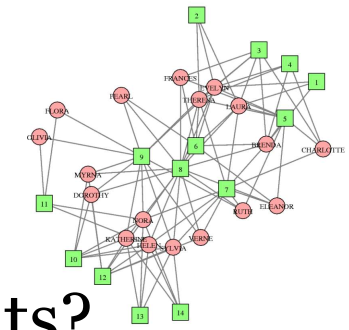
…attending 14 social events

18 women…

1

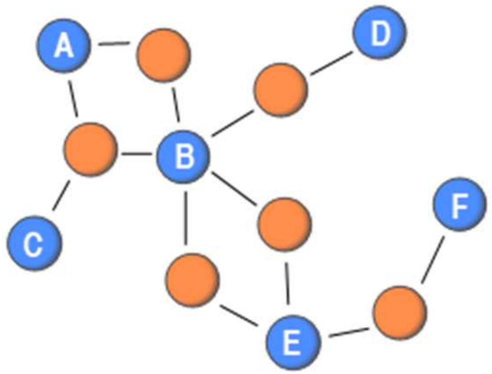
e.g., Helen attends event E10

Texas society ladies



- Projecting 2-mode data to 1-mode
  - How do ladies relate through shared events?
  - How do events relate through shared ladies?
- X: affiliation matrix (rows!=columns)

X

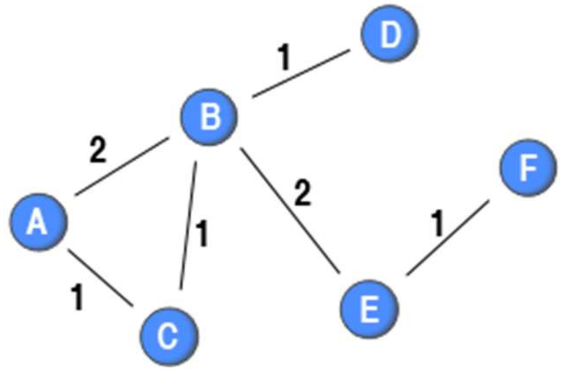


2-mode

Project

(on blue actorset)

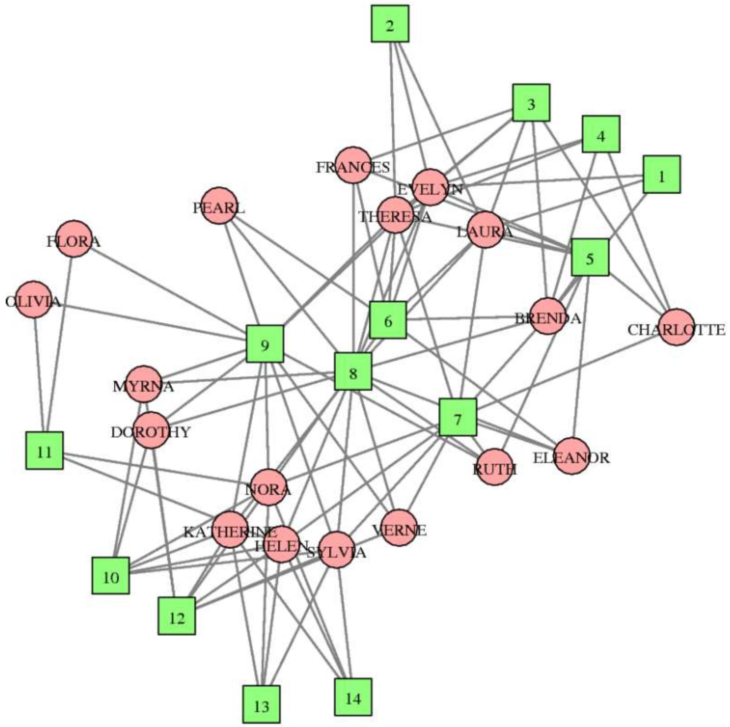
$P = X \cdot X^T$



1-mode

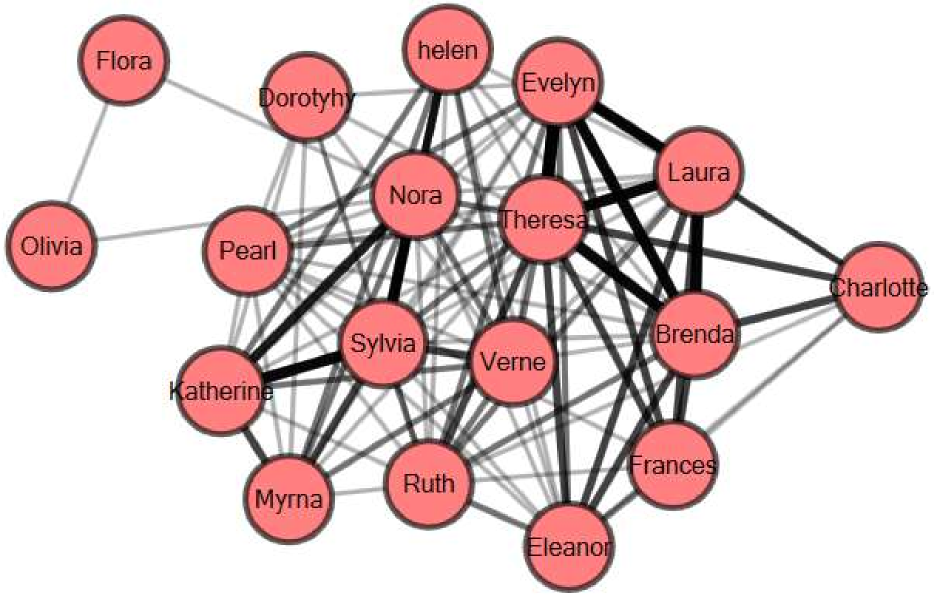
Texas society ladies

2-mode (ladies-to-events)



X

1-mode (ladies-to-ladies)



$P = X \cdot X^T$

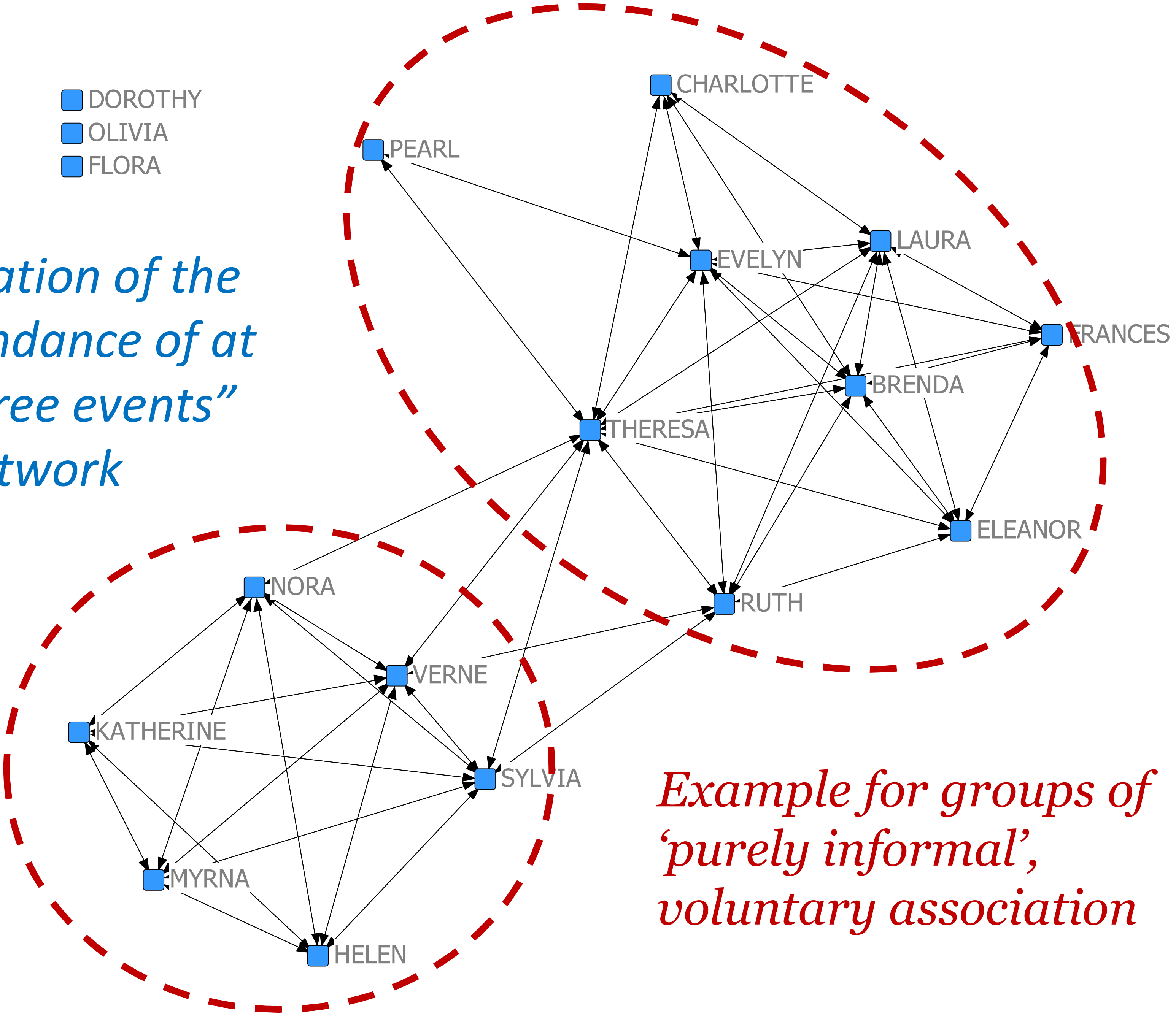
Women-by-women co-attendance valued network

		EV	LA	TH	BR	CH	FR	EL	PE	RU	VE	MY	KA	SY	NO	HE	DO	OL	FL
1	EVELYN	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1
2	LAURA	6	7	6	6	3	4	4	2	3	2	1	1	2	2	2	1	0	0
3	THERESA	7	6	8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1
4	BRENDA	6	6	6	7	4	4	4	2	3	2	1	1	2	2	2	1	0	0
5	CHARLOTTE	3	3	4	4	4	2	2	0	2	1	0	0	1	1	1	0	0	0
6	FRANCES	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	1	0	0
7	ELEANOR	3	4	4	4	2	3	4	2	3	2	1	1	2	2	2	1	0	0
8	PEARL	3	2	3	2	0	2	2	3	2	2	2	2	2	2	1	2	1	1
9	RUTH	3	3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1
10	VERNE	2	2	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1
11	MYRNA	2	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1
12	KATHERINE	2	1	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	1	1
13	SYLVIA	2	2	3	2	1	1	2	2	3	4	4	6	7	6	4	2	1	1
14	NORA	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	1	2	2
15	HELEN	1	2	2	2	1	1	2	1	2	3	3	3	4	4	5	1	1	1
16	DOROTHY	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	2	1	1

e.g., Eleanor attends 3 events that also Ruth attends

"Projection" of two-mode data to the women-mode.

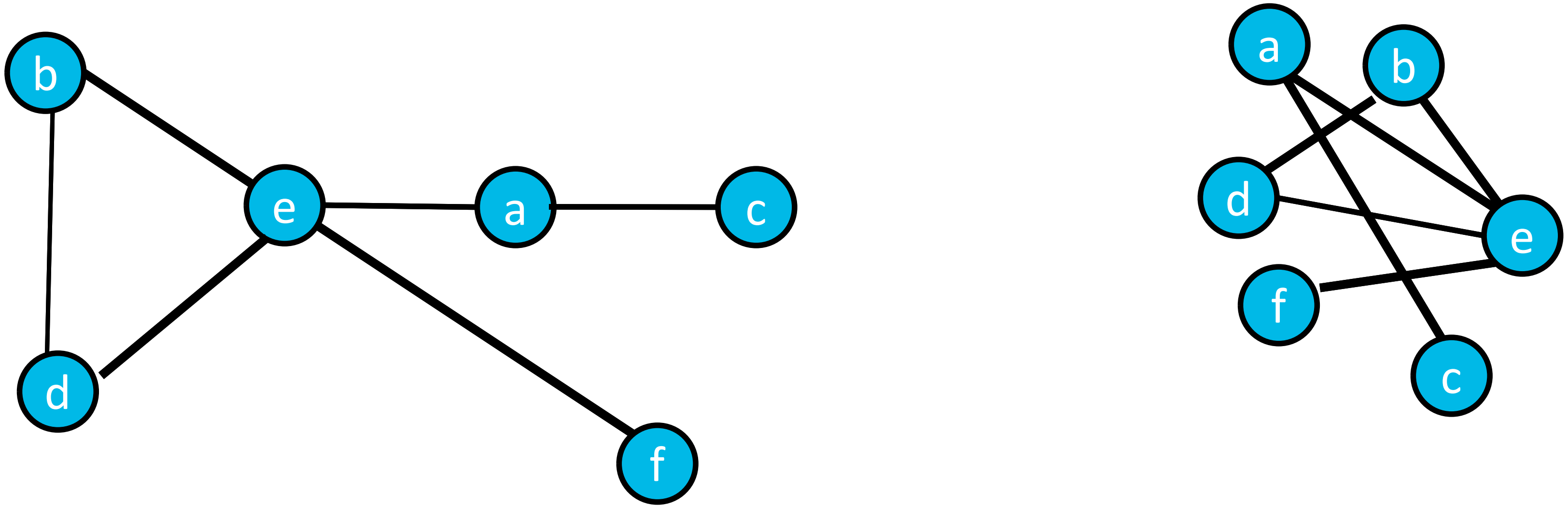
Visualisation of the "co-attendance of at least three events" network



Example for groups of 'purely informal', voluntary association

Visualization principles

What is the difference?



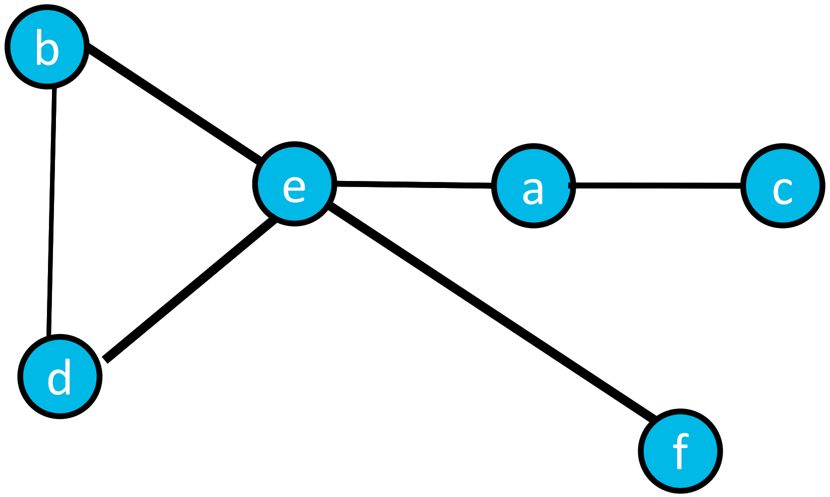
- These graphs are *isomorphic*

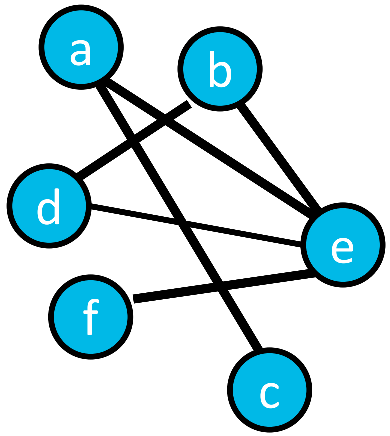
Visualization principles

- Vertex color and line strength
  – Labels on nodes, edges, shapes of lines, nodes
  – Size of nodes (e.g., by degree)
- Layout algorithms
- Options for large networks
- Longitudinal data

Layout principles

- Minimize edge crossings

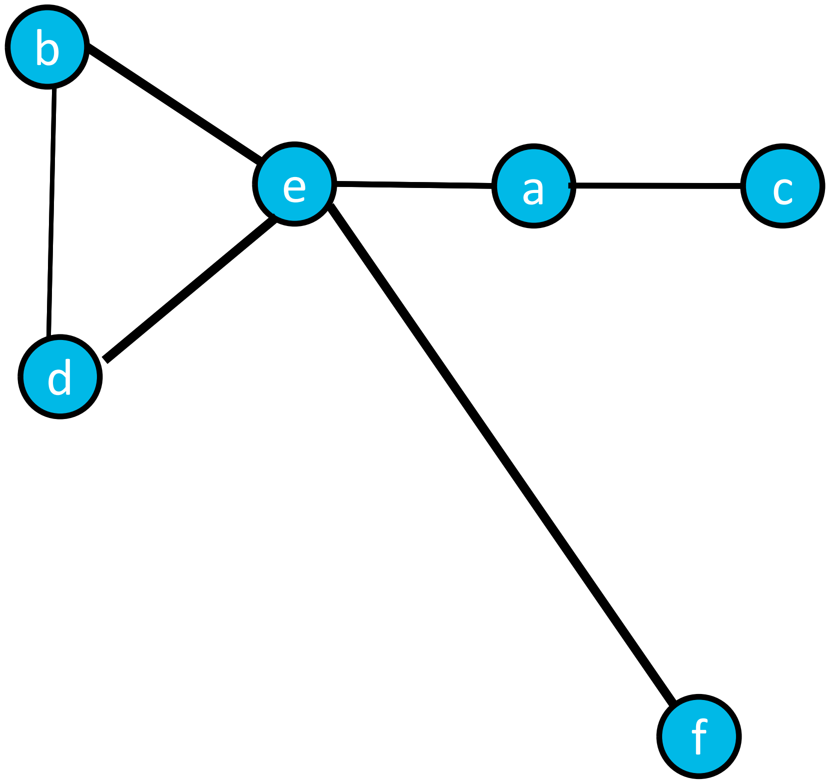


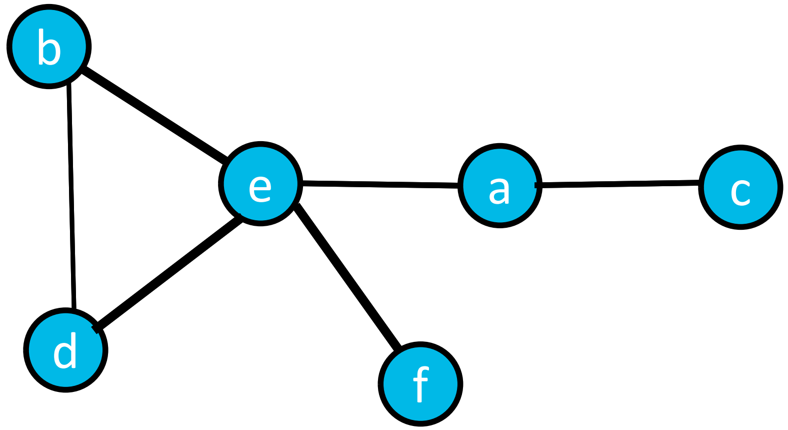


This visualization is more informative

Layout principles

- Minimize edge length variance

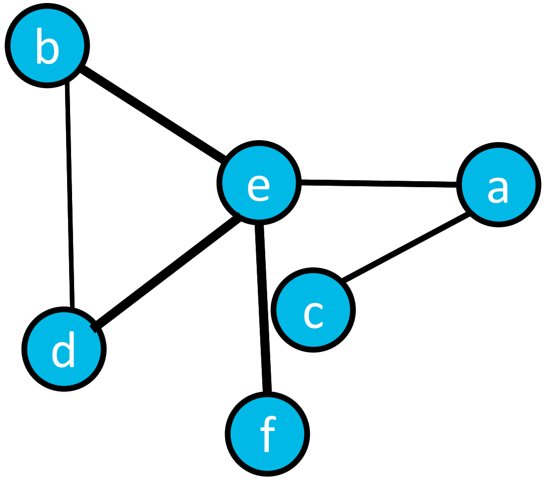


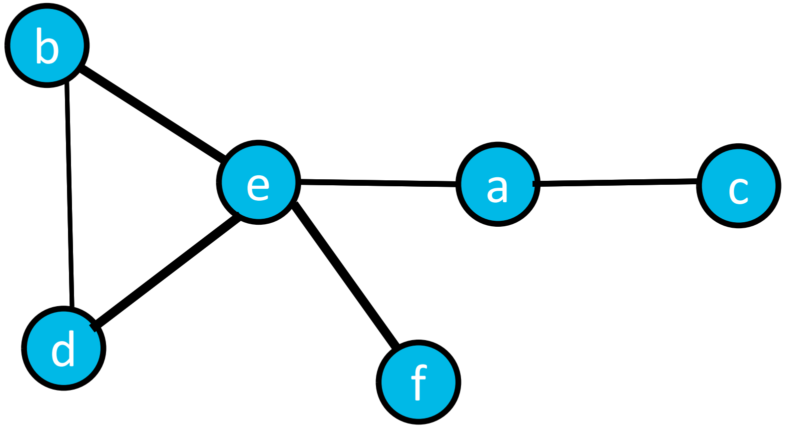


This visualization is more appropriate

Layout principles

- Do not place nodes close to non-incident edges and parts of the network





This visualization is more appropriate

Built-in options for visualization

- No worries: packages have built-in optimizing algorithms for visualization
- Force-directed graph drawing
  - Also known as spring embedders
  - Principle: use a physical analogy to draw graphs, to get adjacent nodes close to each other, and distant nodes to a distant location in space
  - Kamada-Kawai
  - Fruchterman-Reingold
  - Advantages: aesthetically pleasing, exhibit symmetries, minimal amount of crossings
  - Disadvantage: local optimizers, iterated algorithms starting from a random visualization, hence DIFFERENT results for the same graph

Alternatives to force-directed graph visualization

- Circle layout
  - Ego-networks
  - Small world graphs (see later in course)
  - Close-to-regular graphs
- Tree layout (food webs, close-to-hierarchies)
- Spatial networks

Figure 1. *Russian trade routes in the 12th - 13th centuries.*

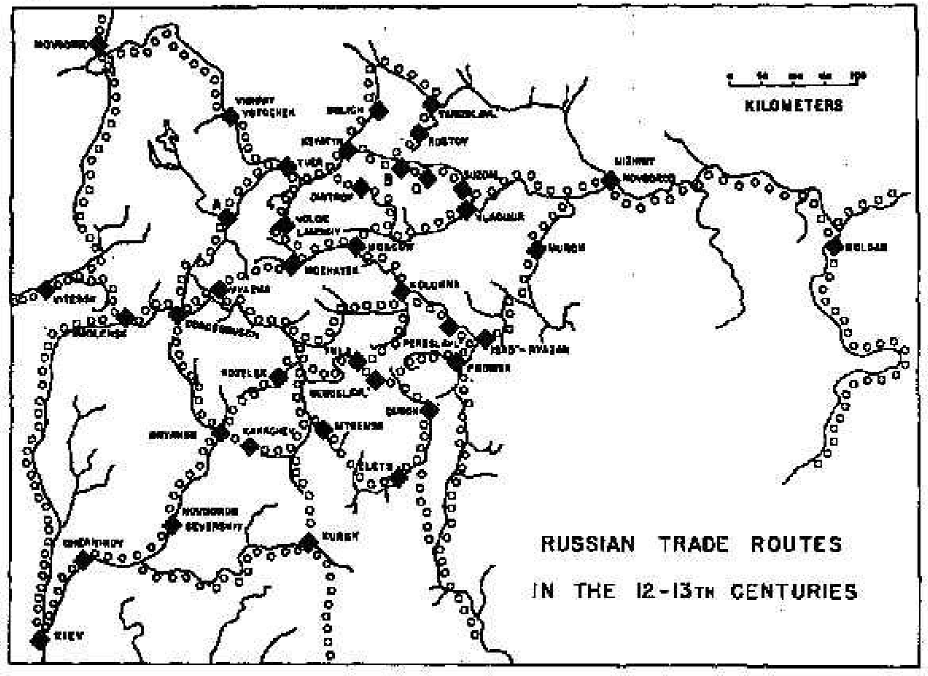
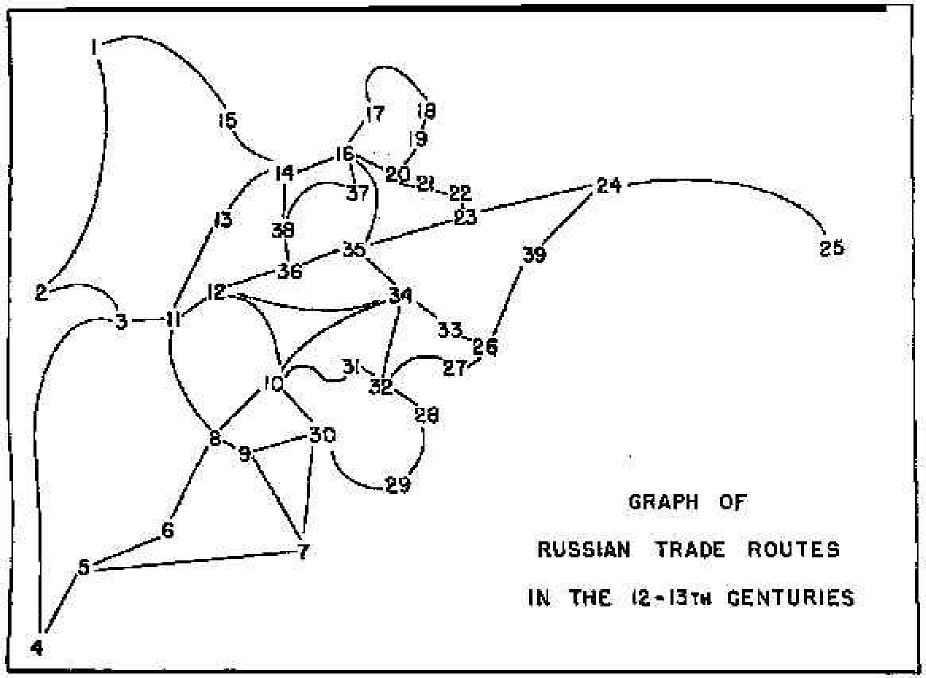


Figure 2. *Graph of Russian trade routes in the 12th - 13th centuries.*



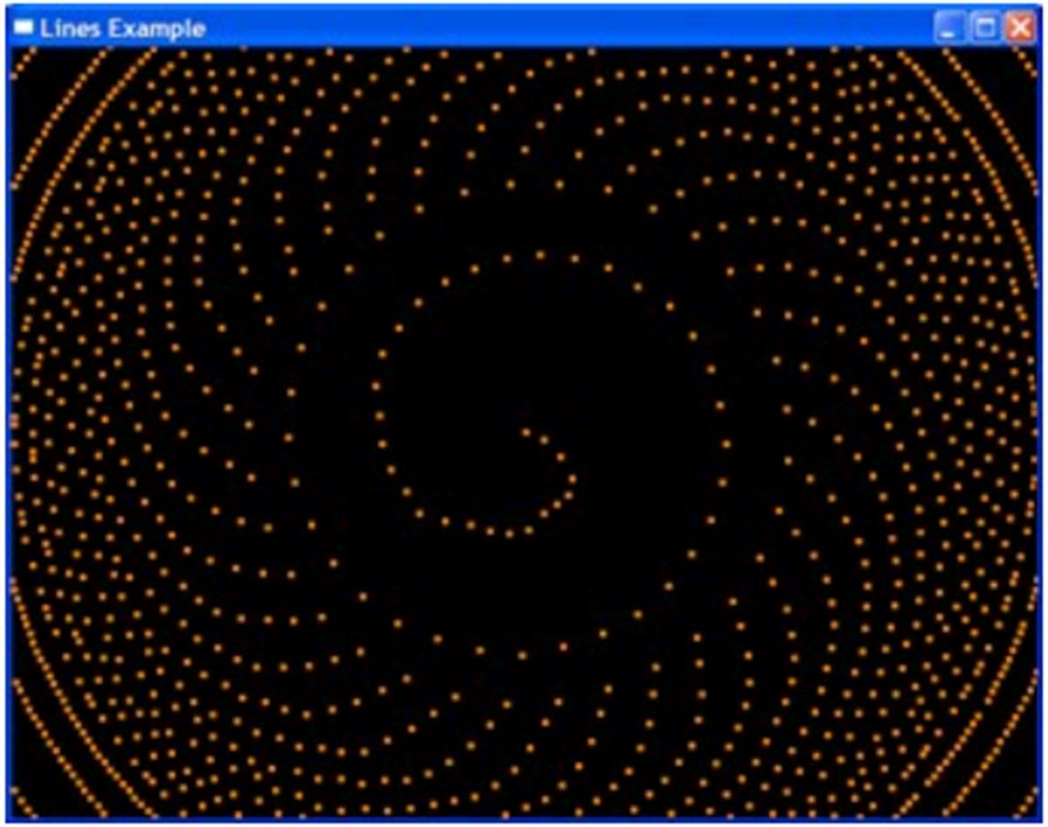
Large networks visualization

- Small networks: small potential misguidance by visualization, large networks: large potential misguidance
- For display of all nodes and ties: impossible
  - Not enough pixels on your screen!
    - 1920 x 1200 = 2,304,000 pixels…
- Massy hairballs even for sparse networks

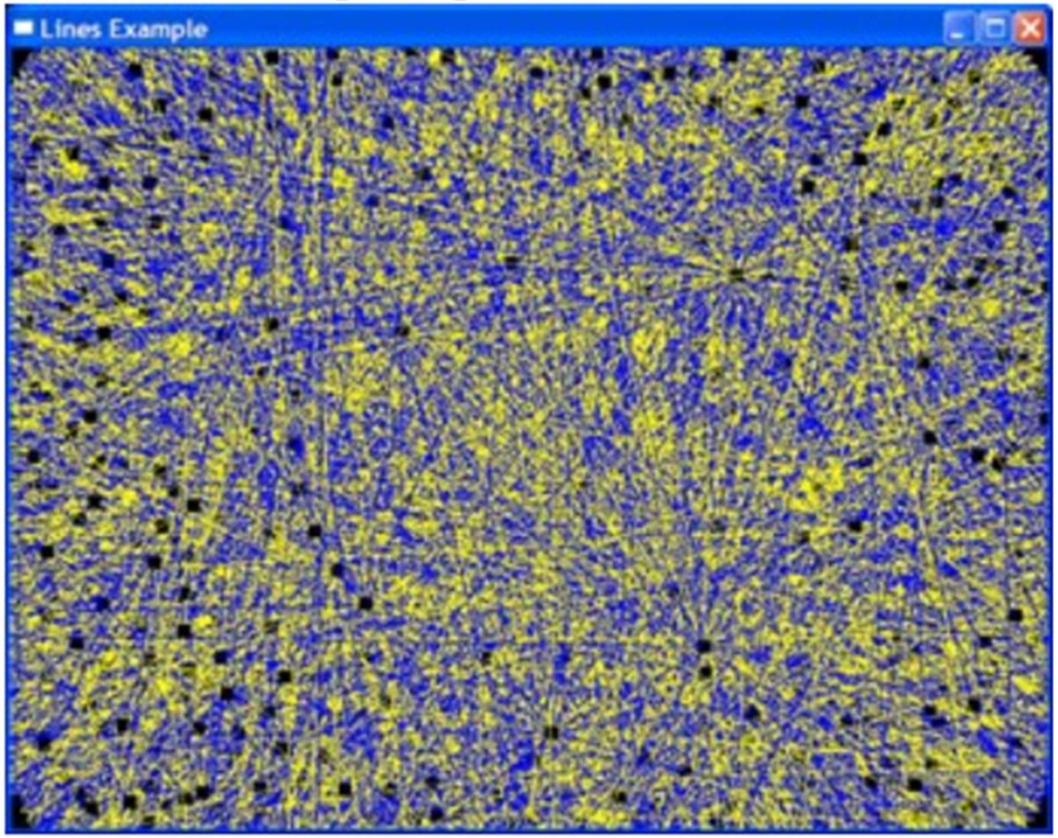
Large networks: congested display

Example by Adamic, L. et al.:

Before drawing edges



After drawing edges



Large networks visualization principles

- Understand the data first and its peculiarities
- Get a first glimpse
- Compress information and group data as much as you can
- Erase non-data, redundant data, unimportant part of data from the visualization
- Revise and repeat

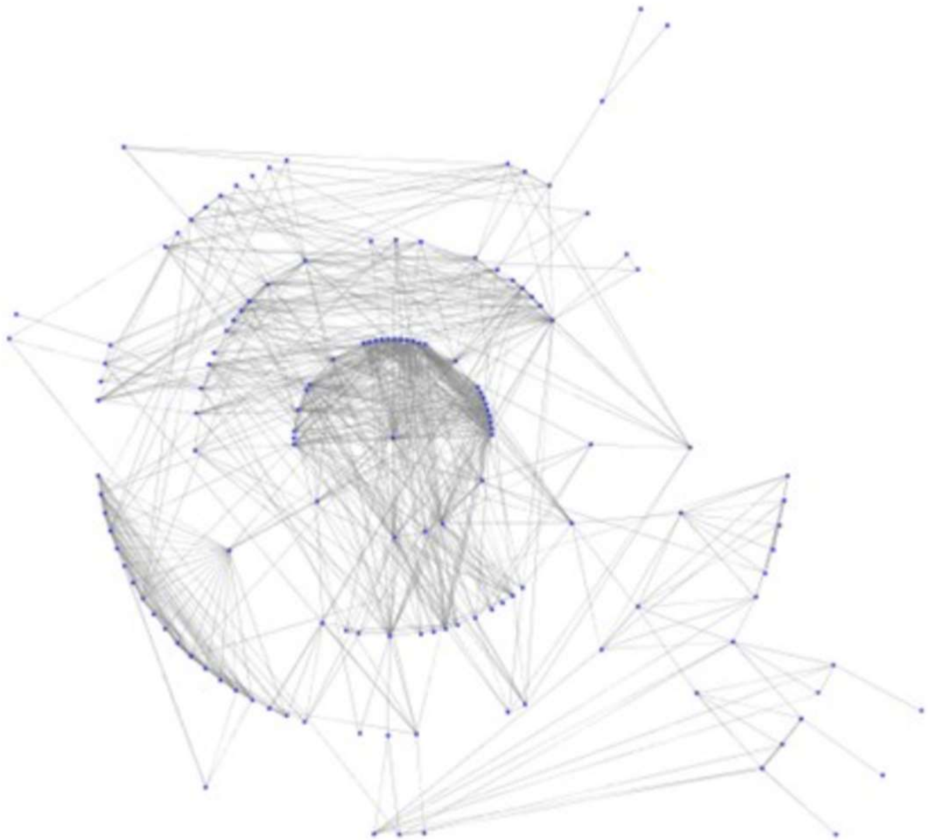
- Viewing only a subset of the network and highlighting key node attributes is often more helpful than a full representation
- Think of visualization as *model of reality to assist understanding*

Radial layout

A nice ”onion” layout for larger networks

- Start with a central node
- Position adjacent nodes in an inner circle around
- Position nodes at distance two in a next circle

Example by Adamic, L. et al.:



Longitudinal networks

- Understand the data first and its peculiarities
- Composition changes, boundary changes, missingness
- Typical aim: compare between times (waves) or display evolution / change
- Fix node positions for a better comparison
  - But where?
  - Optimized for wave 1 (might look ugly for later waves)
  - Better: use some kind of joint optimization