Social Network Analysis

Day 3

SNA Data Collection

Cohesion, Subgroups & Communities

Hypothesis Testing, Inferential Network Models

Collection, Ethics & Entry

Steps to a SNA study

1. Identify the population
   - Bounding, sampling, gaining access

2. Determine the data sources
   - Archival, interviews, observations, surveys

3. Collect the data
   - Survey design

1. Identify the Population: Bounding the Study

- Extremely vexing to beginners and outsiders
  - Network concept would seem to argue against boundaries
- Empirical research makes clear we are all connected
  - Even if distant links don't matter, some people in the sample will be at the edge, no matter where we cut it
- One key is to isolate when bounding matters
  - Yes: Interpersonal influence studies
  - No: Selection studies

Types of Boundaries

- Attribute-based
  - Top management team at Enron
  - Drug injectors in Hartford
- Relation-based
  - Snowballing out from seed sample until few or no new names (i.e., exhaust current component)
- Mixed criteria
  - Sexual ties among residents of Nang Rong
- Theoretical criteria

boundary specification…

What *is the theoretically relevant population?*

Networks are (generally) treated as bounded systems, what constitutes your bound?

	Local	Global
"Realist" (Boundary from **actors**' Point of view)	Everyone connected to ego in the relevant manner (all friends, all sex partners)	All relations relevant to social action ("adolescent peer network" or "Community Health Leaders")
Nominalist (Boundary from **researchers**' point of view)	Relations defined by a name-generator, typically limited in number ("5 closest friends")	Relations within a particular setting ("School friends" or "Physicians serving this hospital")

Most of the time….these boundaries are porous

boundary specification…

In practice:

a) set a pragmatic bound that captures the bulk of theoretically relevant data
b) Collect data on boundary crossing.
    a) You might ask “friends in this neighborhood” but also “Other close friends?”
    b) Don’t limit nominations to current setting, but only trace within the bounds.

Good prior research, ethnography, informants, etc. should be used to identify the bounds as best as possible, but these sorts of data allow one to at least control for out-of-sample effects in models.

For adaptive sampling, such as link-trace designs, you might use a capture/recapture rule to figure out if you’ve saturated your population. Once you stop receiving new names…you’ve finished.

--but, if you jump to a new population…this can be hard to discern.

boundary specification…

1. The level of analysis implies a perspective on sampling:
    1. Local → random probability sampling
    2. Adaptive → Link trace, RDS
    3. Complete → Census

    These are not as dissimilar as they may appear:

    a) Local nets imply global connectivity:
        a) Every ego-network is a sample from the population-level global network, and thus should be consistent with a constrained range of global networks.
        b) If you have a clustered setting, many alters in a local network may overlap, making partial connectivity information possible.
        c) For attribute mixing (proportion of whites with black friends, low BMI with high, users with non-usres, etc.), ego-network data is sufficient to draw global inference

Social Network Data

Research Design: Network Sample

Data collection strategy

	Nominalist (researcher pov)	**Realist (natural groups)**
Local	• Probability samples • Clinical samples • Extracted from complete settings	• Family interviews • Neighbors • Workplace samples
Adaptive	• Fixed diameter chain from qualifying seed(s)	• Unlimited diameter chain on qualifying relation
Complete	• Census within a fixed setting (hospital, school, etc.)	• Only practical for real groups (“Duke Faculty” “Crip”). Get list from informant & enumerate.

(The column distinction is squishy…)

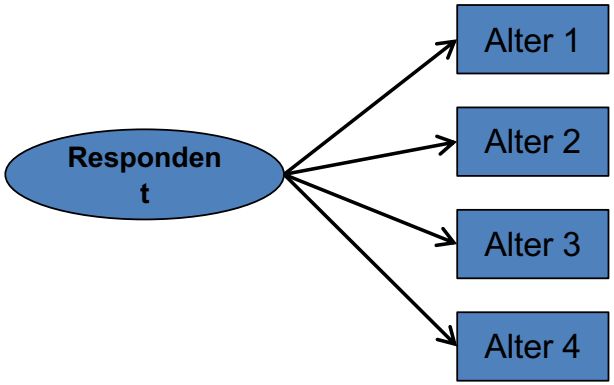
Social Network Data

Research Design: Network Sample

1. Ego Network Sampling

- Most similar to standard social survey:
  - Easily sampled (as any other survey implementation)

- All information comes from the respondent, so very subject to personal projection.



- Ask ego to report on characteristics of alter
  For *k* alters and *q* attributes → adding *kq* questions
  i.e. 5 friends with 10 behaviors adds 50 questions to the survey!

- Ask ego to report on relations amongst alters.
  For *k* alters and *j* relational features → j(k(k-1)/2) questions
  i.e. 5 friends and 2 relation question is 20 questions: 2*((5*4)/2)

Social Network Data

Research Design: Network Sample

2. Snowball and "link trace" designs

Basic idea is to use "adaptive sampling" – start with (a) seed node(s), identify the network partners, and then interview them.

Earliest "snowball" samples are of this type. Most recent work is "respondent driven sampling. (*RDS*)"

-- If done systematically, some inference elements are knowable. Else, you have to try and disentangle the sampling process from the real structure

Social Network Data

Research Design: Network Sample

3. Global network samples: Population Census

- Key issue is to enumerate the population & collect relational information on all.

  - If dynamic, this can make implementation difficult
  - Tends to force case-study style designs (highly clustered settings)
  - Contrast N of networks with N of respondents
  - Because behavior is self-reported (rather than alter reported), adding network questions to a census-based survey is low cost.
    - If you are doing a census anyway….then good to add network questions.

Sampling

- Sampling is not a problem for ego networks

- Sampling for complete networks is in its infancy

Gaining Access

- A little harder than for ordinary studies
  - Strong preference for complete data
  - Respondent fears
  - Length of interview
- Quid pro quo helps but muddies the ethical waters

Step 2: Determine Data Sources

- Archival data
- Interviews
- Observations
- Surveys

Step 3: Collect the Data

- What questions to ask?
  - How many questions to ask
    - Depends on style (roster v. recall)
- How to format your survey?

What Questions to Ask?

- **IT DEPENDS!!!**
  - A relation is just a variable. “Giving advice” is to network analysis what “attitude toward gun-control” is to survey research.
  - It is the researcher who defines the relations of interest. What’s relevant for the phenomena in question?
    - HIV diffusion: sexual ties and needle-sharing are directly involved, other ties like acquaintanceship can potentially turn into sex and sharing ties

what question to ask?

What *information* do you want to collect?

This is ultimately a theory question about how you think the social network matters and what social or biological mechanisms matter for the outcome of interest. This is driven by thinking through:

Health Outcome → Mechanism → Relation(s)

Examples:
Sometimes the relations are clear:

STD/HIV → Contagion-carrying contact → Sex, Drug sharing, etc.

Sometimes not so much:

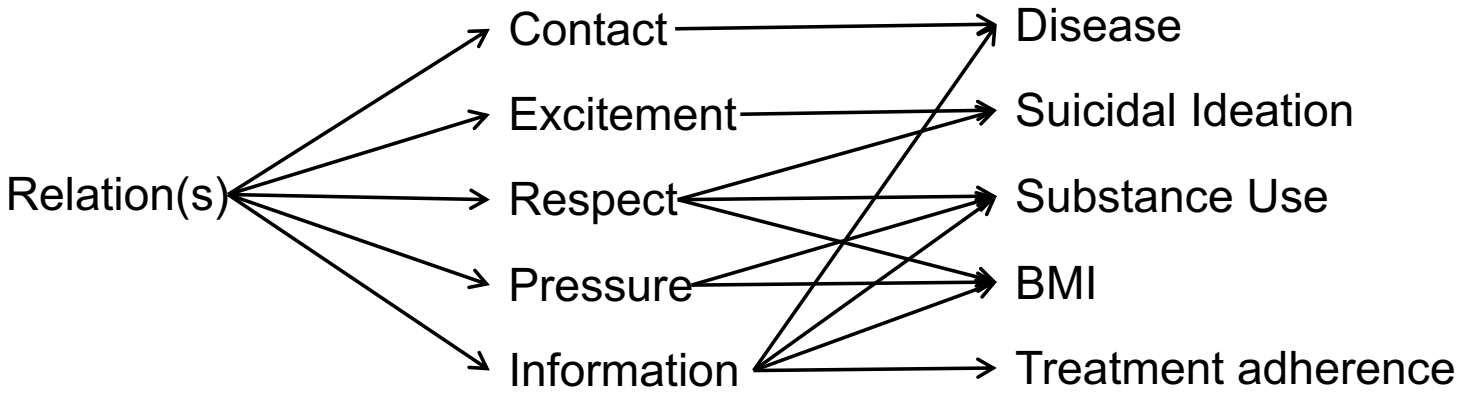
Health Behavior → Information flow → Discussion networks
Health Behavior → Social Conformity Pressure → Admiration nets
Health Behavior → opportunities → Unsupervised interaction

what question to ask?

What *information* do you want to collect?

Sometimes the outcome is deliberately unspecified, as when you are collecting data for a large common use projects (GSS, Add Health, NHRS).

Then the design is effectively reversed: What relations capture the most (general? comprehensive? efficacious? Reliable?) social mechanisms that will be of broad interest?



Social mechanism ambiguity allows broad use, which favors relations that tend to be general. This, of course, makes crisp causal associations more difficult.

what question to ask?

What *information* do you want to collect?

Health Outcome → Mechanism → Relation(s)

Relations themselves are often multi-dimensional…do these matter for your question?

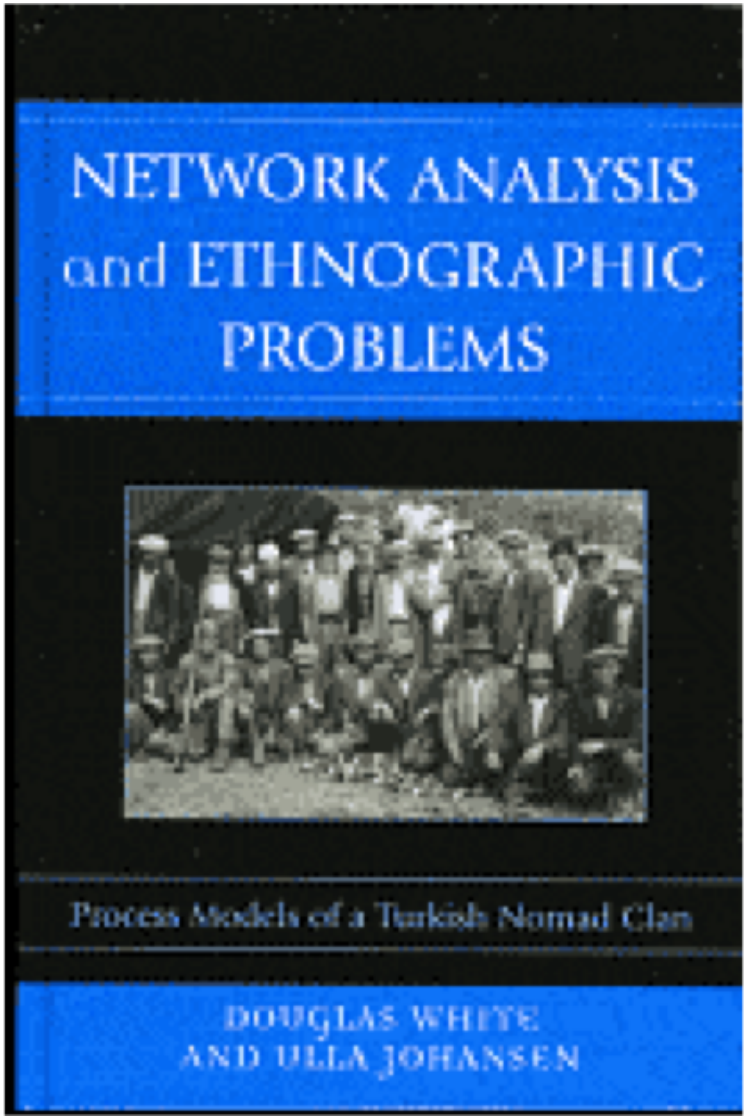
- Perception vs. interaction?
  "who do you like?" ←→ "who do you talk with?"
- Intensity?
  "How often …", "how much…"
  strong vs. weak
- Dynamics?
  Starting & ending dates, everyday contact or sporadic?

Ethnographic Sandwich

- Ethnography at front end helps to …
  - Select the right questions to ask
  - Word the questions appropriately
  - Create enough trust to get the questions answered
- Ethnography at the back end helps to …
  - Interpret the results
  - Can sometimes use resps as collaborators

A Public Service Announcement

- Douglas White has a book about the intermingling of Ethnography and Network Analysis
  - It's a couple years old
- Based on reputation, I expect it is very good, so you might consider looking at this if you are particularly interested in the subject and problem.
- https://goo.gl/eqnJkJ



Surveys

Survey Elements

a) Informed consent
    a) It is important to let people know that their identities matter: network data are confidential but (at least in the construction) not anonymous.

b) Name Generator Questions
    a) General term for what relation you are trying to tap.
    b) Many extant name generators out there…most evidence suggests that people are very sensitive to the questions asked.
        a) If you ask multiple relations, be clear whether it is OK to repeat names!

c) Response Format
    a) Open List → number of lines suggests "right" answer
    b) Check off/select → very simple on/off, might result in over-estimates
    c) Limit choice → limiting choice limits degree which affects *every* network statistics.
    d) Rank/Rate → asking people to rank each other is difficult (and can backfire!)
    e) If multiple name generators – grid or separate questions?

Surveys

If you use surveys to collect data, some general rules of thumb:

a) Network data collection can be time consuming.

If interests are in network-level structure effects, it is better to have *breadth* over *depth*. Having detailed information on <50% of the sample will make it very difficult to draw conclusions about the general network structure.

If interest is in detail interpersonal information – social support for example – detailed information on one or two key ties might be more important.

Survey time is the crucial resource: never enough to ask everything you want.

b) Question format:

- If you ask people to *recall* names (an open list format), fatigue will result in under-reporting
- If you ask people to check off names from a full list, you can often get over-reporting

c) It is common to limit people to ~5 nominations. This will bias network stats for stars, but is sometimes the best choice to avoid fatigue.

Survey Design Issue

- Paper or Plastic?
- Close-ended (Roster) vs. Open-ended
- Repeated Roster vs. MultiGrid
- Tick vs. Rate

Paper or Plastic?

- Paper medium
  - Reliable
  - Reassuring to respondents
  - Errors in data entry
  - Data entry is time-consuming
- Electronic
  - Span distances, time zones
  - Harder to lose
  - Fewer data handling errors
  - Lower response rate
  - Emailed documents vs survey instruments

Closed-Ended vs Open-Ended

Roster of names or just blank lines?

- Closed-ended (aided)
  - Requires bounded list
  - Can be impractical for large networks
- Open-ended (unaided)
  - Subject to recall errors
  - Can limit number of choices made (more effort, limited space)

Name	Q1. Heard of them
Allata, Joan	☐
Baer, Justin	☐
Baker, Ted	☐
Berenwitz, Rick	☐
Branzei, Oana	☐
Brooks, Scott	☐
Brower, Ralph	☐

If you wanted to get something done on behalf of a customer who would you contact? *(write as many names as you like in the spaces provided)*

_____ _____

_____ _____

_____ _____

_____ _____

Hybrid Questionnaire

1. If you wanted to get something improved or done on behalf of a customer who would you contact?

Name	(index no.)
Denny Terio	(169)
Eric Estrada	(27)
	( )
	( )

2. If you wanted to get a true reading on where [company name] was headed as an organization, who would you talk to?

	( )
	( )

Paper version uses separate booklet containing name directory

Web version uses drop-down menus

Q1. Please indicate which of the following you had met or been aware of before coming to this workshop.

Allata, Joan ☐
Baer, Justin ☐
Baker, Ted ☐
.... ☐

Q2. Check of f the names of the people you know. By "know" I mean that you have spoken to each ...

Allata, Joan ☐
Baer, Justin ☐
Baker, Ted ☐
.... ☐

Q1. Using the checkboxes below, please indicate **who you have heard of or know about** among the participants of the workshop.

Q2. Check off the names of the **people you know**. By 'know' I mean that you can attach a name to a face, you have spoken to each other at least once, and the other person is also likely to put you down.

Q3. Check off the names of people you **have worked with** on a paper or other academic/administrative project.

Q4. Check off the the names of a selected set of people whom you don't know but **would like to know**, based on things you've heard, or their interests, etc.

Name	Q1. Heard of them	Q2. Know them	Q3. Worked with	Q4. Want to know
Allata, Joan	☐	☐	☐	☐
Baer, Justin	☐	☐	☐	☐
Baker, Ted	☐	☐	☐	☐
Bercuwitz, Rick	☐	☐	☐	☐
Branzei, Oana	☐	☐	☐	☐
Brooks, Scott	☐	☐	☐	☐
Brower, Ralph	☐	☐	☐	☐

Tick or Rate?

- Ask respondent for yes/no decisions or quantitative assessment?
  - Yes/no are cognitively easier on respondent (therefore reliable, believable),
  - Yes/no *much* faster to administer
  - But yes/no provides no discrimination among levels

- A series of binaries can replace one quant rating:
  - Instead of “How often do you see each person?”
  - 1 = once a year; 2 = once a month; 3 = once a week; etc.
- Use three questions (in this order):
  - Who do you see at least once a year?
  - Who do you see at least once a month?
  - Who do you see at least once a week?

Question Wording Issues

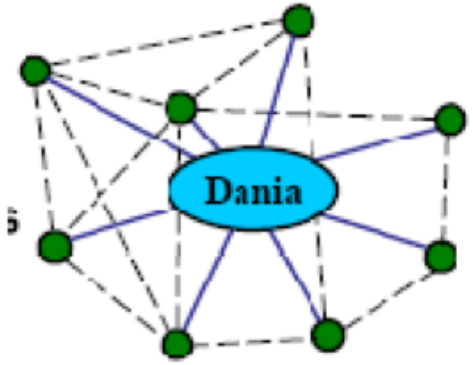
- “Friendship” does not mean the same thing to everyone
  - Especially across national cultures
- Some helpful practices
  - Use one word label plus two or three sentence description, plus have full paragraph detailed explanation available
  - Use homogeneous samples

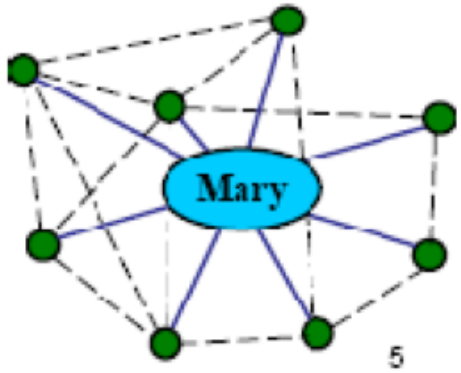
Survey Construction Strategies

- Ego Net
- Row-based (for undirected relations)
- Row and Column-based (for directed relations)
- Matrix based (Krackhardt CSS)

Ego Networks

- (Random) sample of nodes
  - Each sampled node called an "ego"
- Each is asked for set of contacts called "alters"
- Each is asked about attributes of self, and alters
- Ego also asked (usually) about ties among alters
- Connections between ego's or between alters of different egos are not recorded
  - Each ego is a world in itself





Row-Based

- Each informant questionnaire corresponds to one row in the network adjacency matrix
- Issues of comparability across respondents
- For logically undirected relations, can deal with accidental asymmetry and missing respondents via symmetrization
  - Intersection rule: $X_{ij} = 1$ if $X_{ij} = 1$ and $X_{ji} = 1$
  - Union rule: $X_{ij} = 1$ if $X_{ij} = 1$ or $X_{ji} = 1$

Row and Column Based

- Each informant effectively asked to fill out both their row and their column of the adjacency matrix (but actually stored as separate matrices)
  - A_{ij}: Who do you give advice to?
  - B_{ij}: Who do you get advice from?

- Handle asymmetry by creating new matrix $X = A \cap B^T$ (intersection criterion)
  - $X_{ij} = 1$ iff $(A_{ij} = 1)$ AND $(B_{ji} = 1)$
  - i.e., i gives advice to j if i says i gives advice to j and j says they receive advice from I

- Problem with cognitive & affective relations
- Respondent is the expert

Matrix-based

- Krackardt CSS
- Each respondent asked about relations among all pairs of persons in group, not just those involving self
  - Yields network matrix C(k) for each respondent
- Aggregate respondent matrices using choice of rules
  - Local: $X_{ij} = 1$ if $C(i)_{ij}$ and $C(j)_{ij}$
  - Global: $X_{ij} = 1$ if $C(k)_{ij} = 1$ for most k

Krackhardt CSS

Q1. How well the members of each pair know each other:									
	Response scale: Blank = They have never met. 1 = They are merely								
Knowledge	Aaron	Ali	Dan	Dave	David	Ed	George	Greg	Howard
Aaron									
Ali									
Dab									
Dave									
David									
Ed									
George									
Greg									
Howard									

How *Reliable* are SNA data?

- Response bias
- Asymmetry
- Missing data
- Accuracy
- Ethics

Response Bias

- Some respondents positively biased
  - Give big numbers in general when rating strength of tie or frequency
- Row-based approach yields matrices in which each row potentially has different measurement scale
  - Can create asymmetry when none “exists”
- For valued data can normalize by rows
  - Z-scores, euclidean norms, maximum, marginals

Unexpected Asymmetry

- A claims to have sex with B, but B does not claim to have sex with A
  - The relation is logically symmetric, but empirically asymmetric
  - Errors of recall; strategic response
- Sometimes asymmetry is the point
- Logically symmetric data may be symmetrized
  - If either A or B mentions the other, it's a tie
  - Only if each mentions the other is it a tie

Non-symmetric Relations

- Gives advice to
- Can't symmetrize logically non-symmetric relations, except by changing meaning of tie
- Unless you ask question both ways:
  - Who do you give advice to?
  - Who gives advice to you?
- Two estimates of the A→B tie, and two estimates of the A←B tie

Missing Data

- For logically symmetric relations
  - if X_{ij} is missing, substitute X_{ji}
  - If whole row missing, substitute corresponding column
- For logically non-symmetric relations, ask questions both ways (who do you give advice to, who gives advice to you)
  - set $A_{ij} = B_{ji}$
  - i.e., missing row is replaced with column of the inverse relation

What to do about missing data?

Easy:

- Do nothing. If associated error is small ignore it. This is the default, not particularly satisfying.

Harder: Impute ties

- If the relation has known constraints, use those (symmetry, for example)
- If there is a clear association, you can use those to impute values.
- If imputing and can use a randomization routine, do so (akin to multiple imputation routines)
- All ad hoc.

Hardest:

- Model missingness with ERGM/Latent-network models.
  - Build a model for tie formation on observed, include structural missing & impute. Handcock & Gile have new routines for this.
  - Computationally intensive…but analytically not difficult.

Informant Accuracy

- Bernard, Killworth et al compared observed with recalled interaction data
  - Ham radios, deaf TTYs
  - About half of the cells in the adjacency matrix were wrong
- Romney & Faust noted that structural analyses didn't seem so far off
  - Surface structure vs deep structure
- Freeman, Romney & Freeman
  - Respondents biased toward long term patterns

Krackhardt CSS

- Many sources of inaccuracy
  - Recall and exaggeration of ties with high status people
  - Idiosyncratic understanding of the question
- Take “average” of everyone’s perception of given dyad’s relationship
  - Great for deliberately hidden relationships

Dillman Survey Design Considerations

- Network questionnaires can be fun but are usually time-consuming and generate anxiety
- Providing value
- Treating respondent with respect
- Attractive formatting
- Cloaked in authority and importance

Ethical issues

Ethical and Strategic Issues

- What makes network research especially challenging ethically?
- What are the dangers & to whom?
  - In academic setting
  - In management setting
  - In mixed situations
  - In national security setting
- What can we do about it?

Ethical Issues

- Respondents cannot be anonymous
- Non-respondents are still included
- Missing data can be powerful
- Has the potential to be mis-used by Management

The Belmont Report: Guiding Ethical Principles to Social Science Research

Respect for Persons

Autonomy

Voluntariness

Informed Consent

Beneficence

Do not harm

Maximize possible benefits/Minimize Possible Harms

Justice

The risks and benefits of research should be equitably distributed

Questions of Informed Consent and Privacy

Key Components of Informed Consent

Disclosing to potential research subjects information needed to make an informed decision

Facilitating the understanding of what has been disclosed

Promoting the voluntariness of the decision about whether or not to participate in the research.

Risks in Social Network Studies

In most social network research, the chief risk to respondents is that of being stigmatized as a result of being identified as belonging to a stigmatized category or group (e.g., sex workers, drug addicts), or from adverse consequences resulting from revealing an individual's role or position in a social setting (e.g., discovering you are the least liked individual in your organization).

Social network research shares these risks with other forms of survey-based research that examine the impact of one's social environment on phenomena such as risk taking, mental health, and attitudes towards medical providers.

However, there are some unique sources of risk.

Potential Risks Associated with Relational Data

Outing People

Minor: Mom Finds Out Mike Smokes

Major: Wife Finds Out that Her Husband Has Been Cheatir

Legal Risks

If you trace a relationship between an adult and a child that
would be treated as contributing to the delinquency of a minor, are you legally obligated to report the relationship?

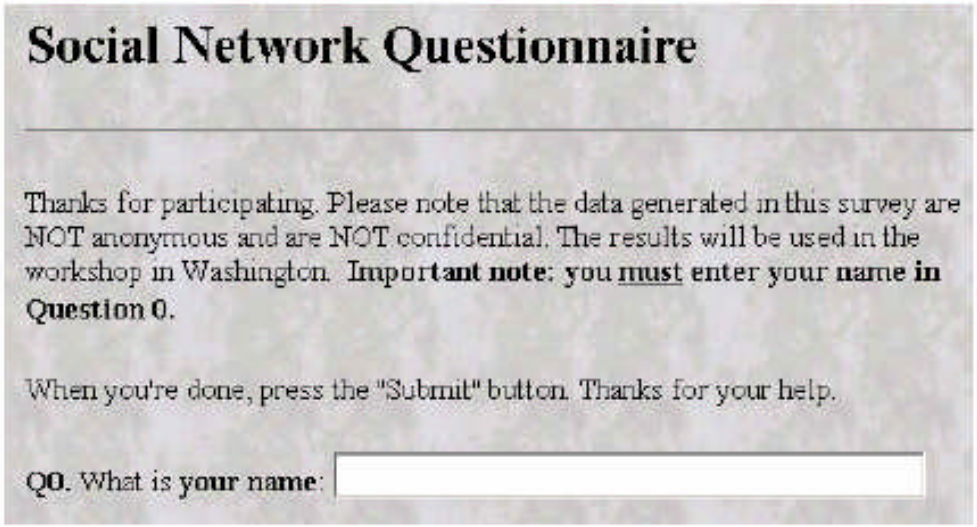
If a known-to-be STD positive person names a partner, do we inform the partner of the respondent's STD status?

Detecting Fraud

Network analyses can reveal inconsistencies that suggest fraud (very high degree, say, or sharing patients in a way that is highly irregular

Confidentiality Reminder

- This is in addition to consent form



Social Network Questionnaire

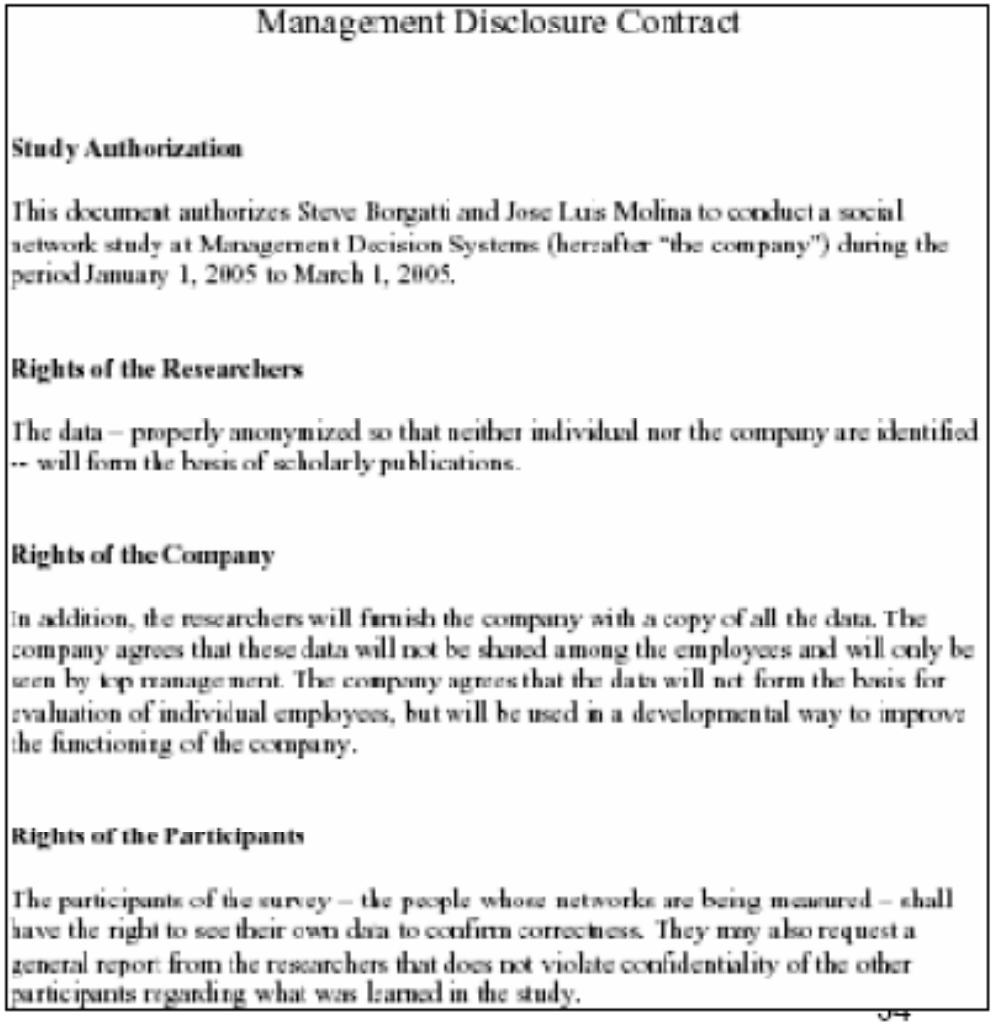
Thanks for participating. Please note that the data generated in this survey are NOT anonymous and are NOT confidential. The results will be used in the workshop in Washington. **Important note: you must enter your name in Question 0.**

When you're done, press the "Submit" button. Thanks for your help.

Q0. What is **your name**:

3-Way Disclosure Contract

- For research done in organizations
- Signed by management, the researchers, and each participant
- Clearly identifies what will be done with the data



Management Disclosure Contract

Study Authorization

This document authorizes Steve Borgatti and Jose Luis Molina to conduct a social network study at Management Decision Systems (hereafter "the company") during the period January 1, 2005 to March 1, 2005.

Rights of the Researchers

The data – properly anonymized so that neither individual nor the company are identified -- will form the basis of scholarly publications.

Rights of the Company

In addition, the researchers will furnish the company with a copy of all the data. The company agrees that these data will not be shared among the employees and will only be seen by top management. The company agrees that the data will not form the basis for evaluation of individual employees, but will be used in a developmental way to improve the functioning of the company.

Rights of the Participants

The participants of the survey – the people whose networks are being measured – shall have the right to see their own data to confirm correctness. They may also request a general report from the researchers that does not violate confidentiality of the other participants regarding what was learned in the study.

Truly Informed Consent Form

Truly Informed Consent Form

Introduction

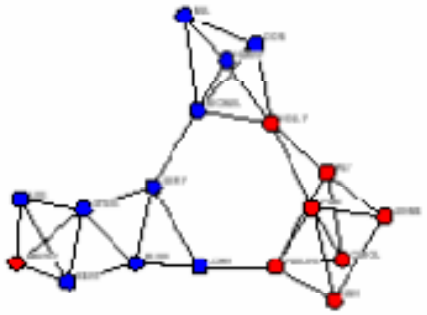
This is a social network study in which we will try to map out the communication network of the organization.

Goals

The academic goal of this study is to understand the factors that determine who talks to whom. We want to understand what factors hinder communication, and which ones facilitate communication. The organization's goal in this study is to improve communication in areas that need it.

Procedures

You will be asked to fill out an online survey about who you interact with regularly, along with background information about yourself, such as training, department you're in, and so on. It should take about 30 minutes to complete. In order to map out who talks to whom, we will need you to give us your name when filling out the survey. Once the data have been collected, we will construct social network maps like this one:



Note that the maps contain each person's name. These maps will be shown to management (specifically, all officers in the organization), but will not be shown to others in the organization. In addition, we will calculate network metrics such as calculating the "degrees of separation" between pairs of people (i.e., the length of the network paths from one person to another).

:

Truly Informed Consent Form

Risks & Costs

Since management will see the results of this study, there is a chance that someone in management could consider your set of communication contacts to be inappropriate for someone in your position, and could think less of you. Please note, however, that the researchers have obtained a signed agreement from management stipulating that the data will be used for improving communication in the company and will not be used in an evaluative way.

Individual Benefits

We will provide you with direct, individualized feedback regarding your location in the social network of the organization.

Withdrawal from the Study

You may choose to stop your participation in this study at any time. If so, you will not appear on any of the social network maps and no metrics will be calculated that involve you. Note that management has agreed that participation in the study is voluntary.

Confidentiality

As explained above, your participation will not be anonymous. In addition, all of top management will be able to see results of the study that include your name. Outside of top management, however, the data will be kept confidential. Any publicly available analyses of these data will not identify any individual by name, nor identify the organization.

Participant's Certification

I have read and I believe I understand this Informed Consent document. I believe I understand the purpose of the research project and what I will be asked to do. I understand that I may stop my participation in this research study at anytime and that I can refuse to answer any question(s). I understand that management and only management will see the results of this research with individuals identified by name.
I hereby give my informed and free consent to be a participant in this study.

Signatures:

Data Agreements

When collecting data establish:

Who owns the data

How will it be collected

Who stores and processes it

How long will identifying information be retained

Who has access to identifying information

The answers to these questions can help in determining whether you believe the study can be conducted in an ethical

Summary

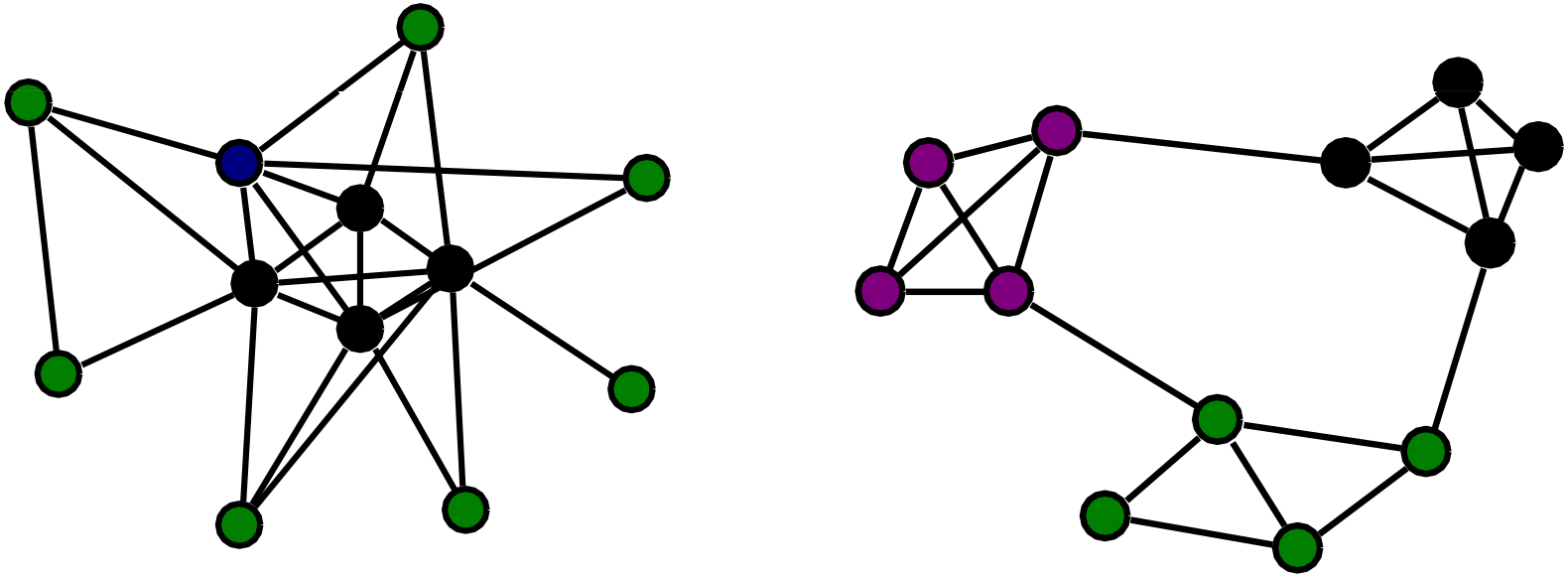
- There are three steps to getting started on a social network study
  - Identify the population
  - Determine data sources
  - Collect data

- In addition there are a number of issues that must be considered such as response bias, missing data, unexpected asymmetry, and ethical considerations

2. Cohesion, Subgroups & Communities

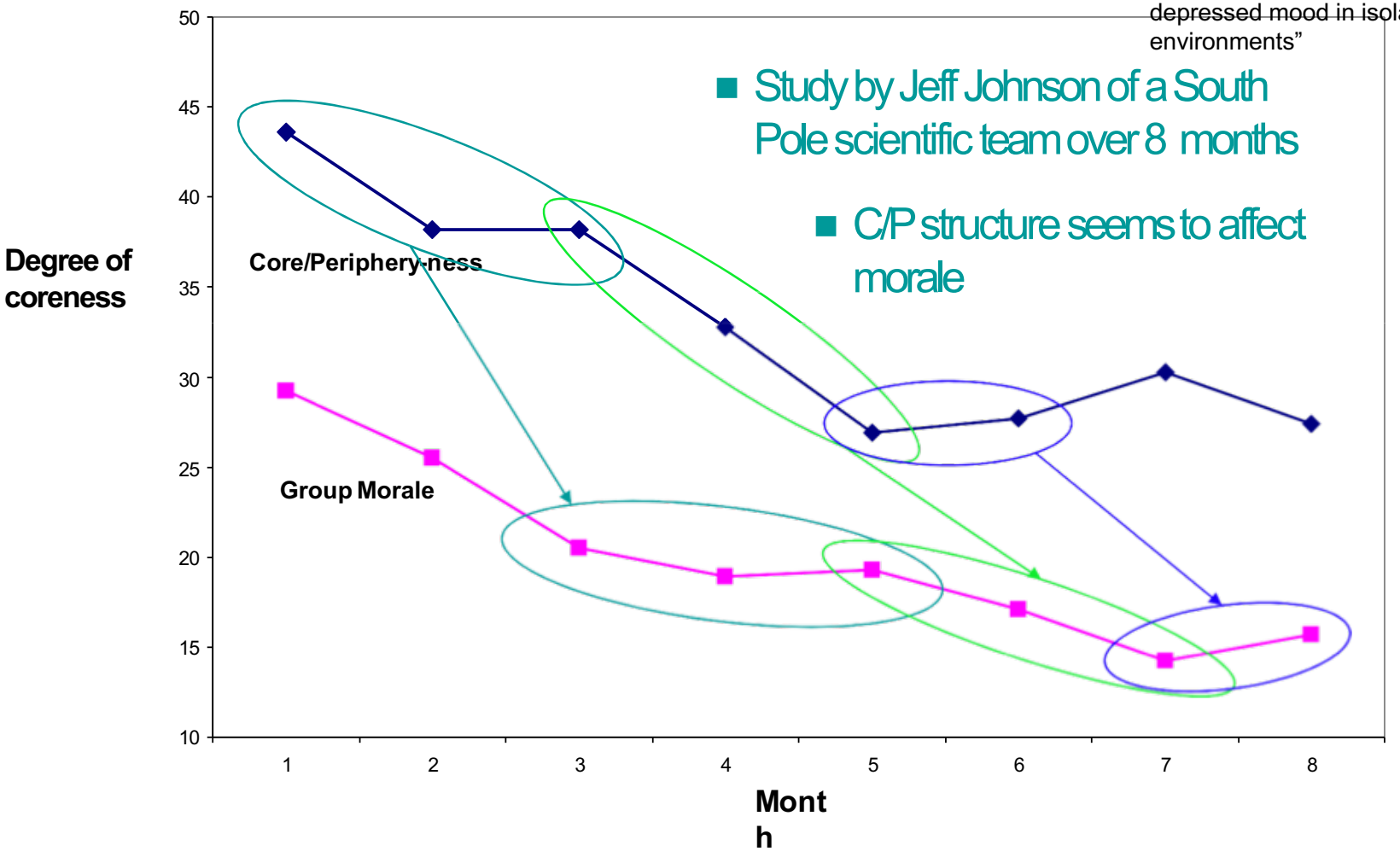
Application

- How do you think network structure interacts with the morale of the group?



Core-Periphery Structures & Morale

Johnson et al. (2003) “Social Roles and the Evolution of Networks in Extreme and Isolated Environments”;
Palinkas et al. (2004) “Social Support and depressed mood in isolated and confined environments”



- peripheral individuals would often develop thyroid problems, which is related to depression;
- globally coherent networks were associated with group consensus

Dyadic & Whole Network Cohesion

- Dyadic cohesion refers to pairwise social closeness
- Whole network measures can be
  - Averages of dyadic cohesion
  - Measures not easily reducible to dyadic measures

- We are going to focus on the whole network parts of cohesion.

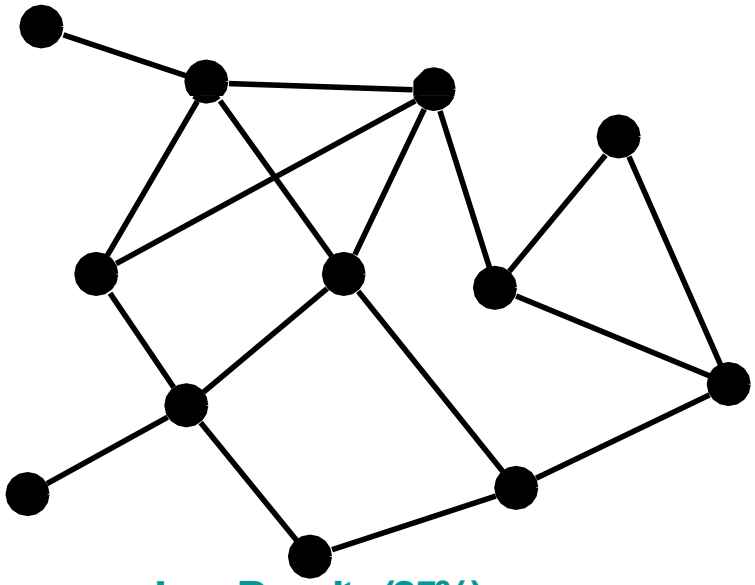
Measures of Group Cohesion

Whole Network Measures

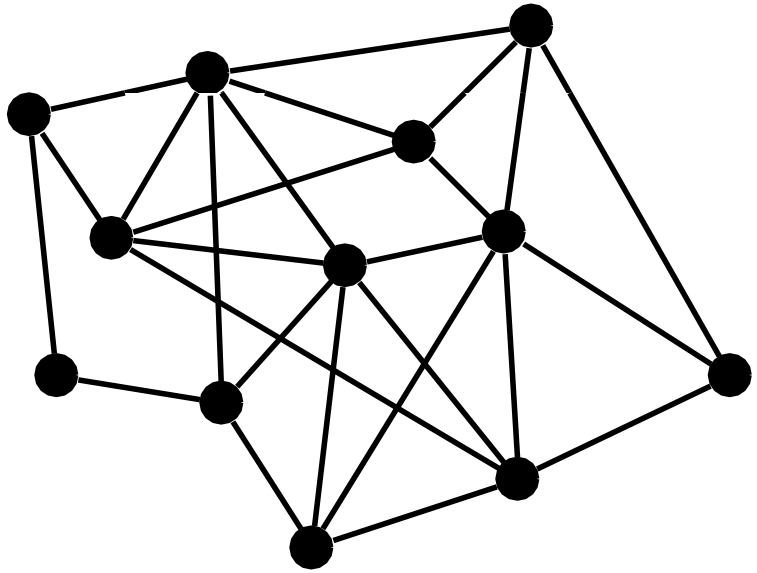
- Density & Average degree
- Average Distance and Diameter
- Component measures (# & Ratio)
- Fragmentation (reachable & distance-weighted)
- Connectivity
- Centralization
- Core/Peripheriness

Density

- Number of ties, expressed as percentage of the number of ordered/unordered pairs

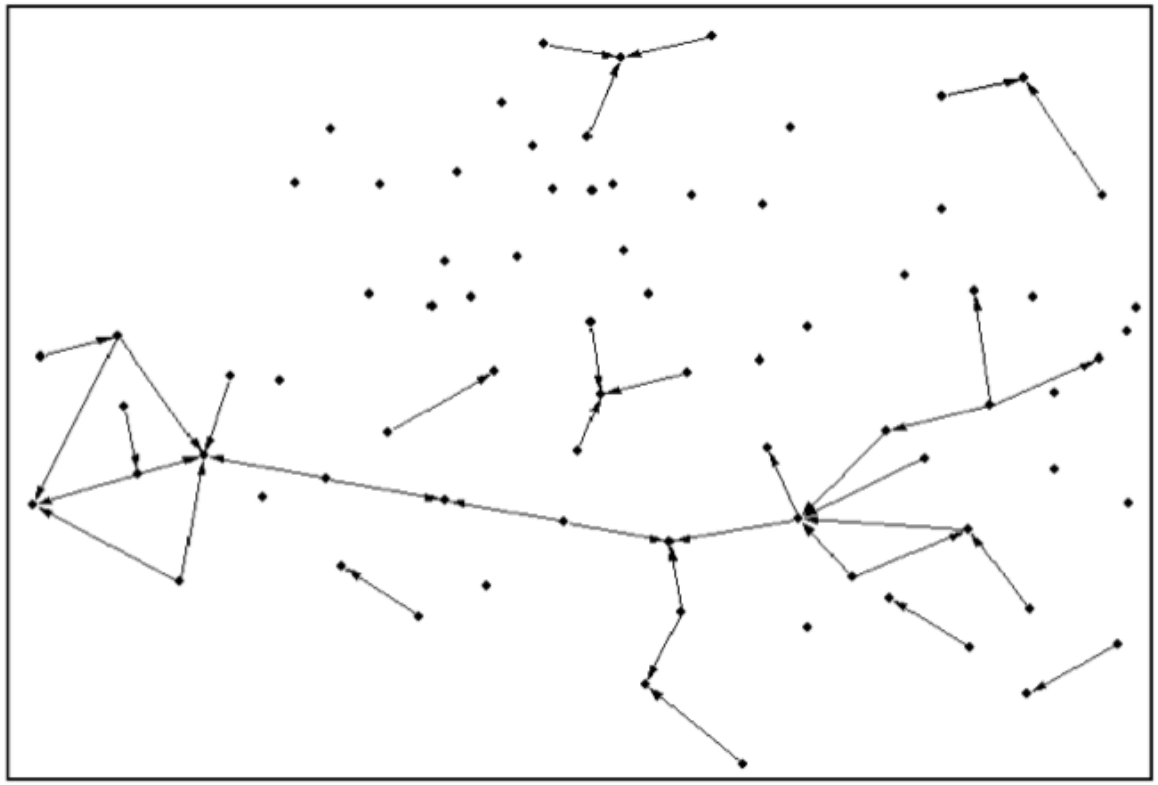


Low Density (25%)
Avg. Dist. = 2.27



High Density (39%)
Avg. Dist. = 1.76

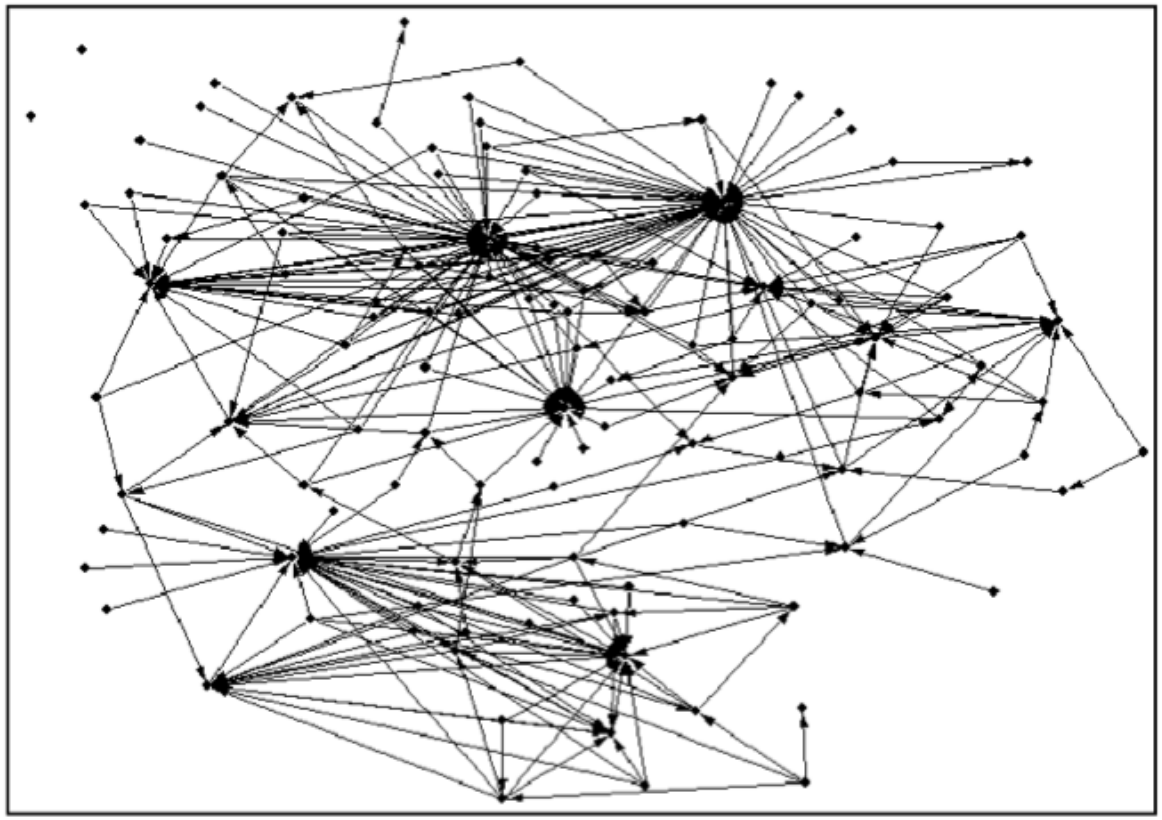
Help With the Rice Harvest



Village
1

Data from Entwistle et al

Help With the Rice Harvest



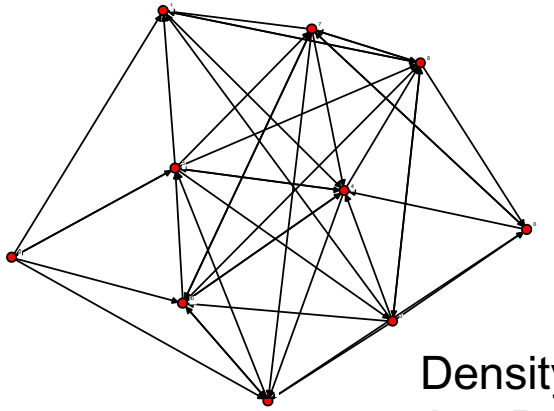
Village 2

Which village is more likely to survive ?

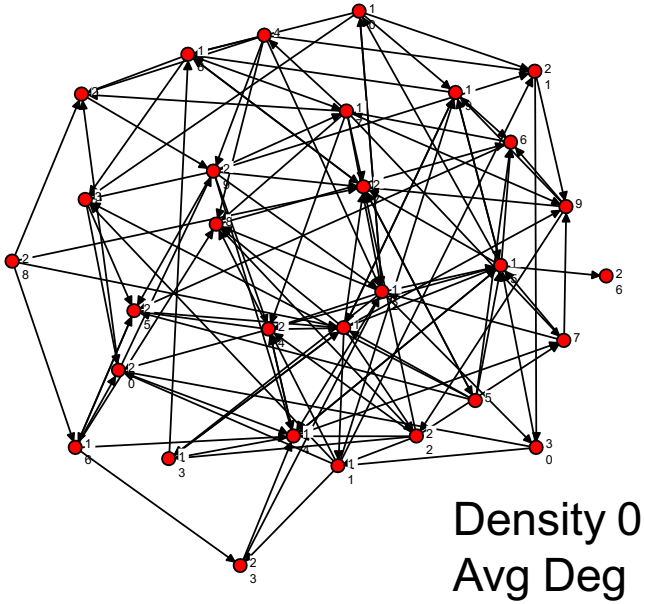
Data from Entwistle et al

Average Degree

- Average number of links per person
- Is same as density*(n-1), where n is size of network
  - Density is just normalized avg degree
  - Often more intuitive than density



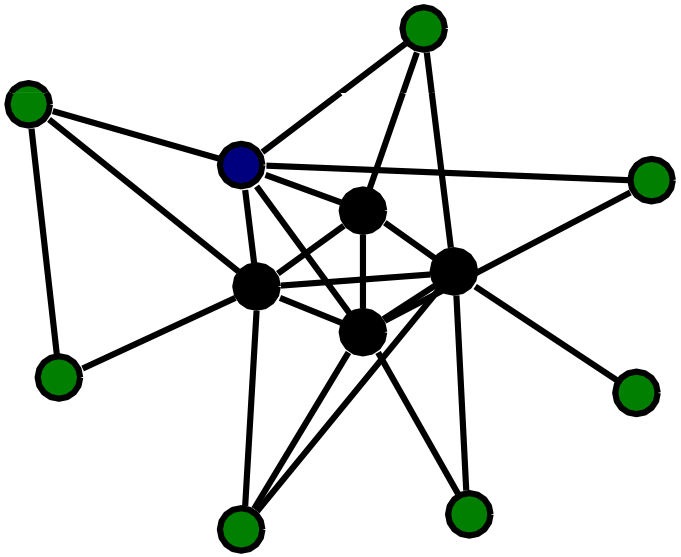
Density 0.47
Avg Deg 4



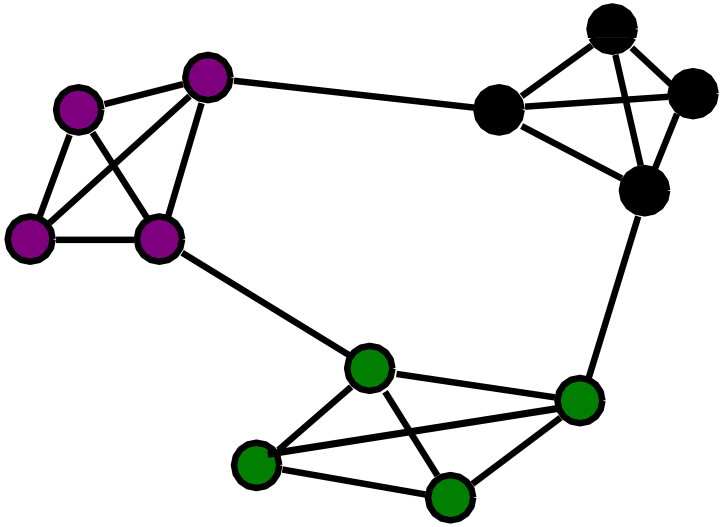
Density 0.14
Avg Deg 4

Average Distance

- Average geodesic distance between all pairs of nodes



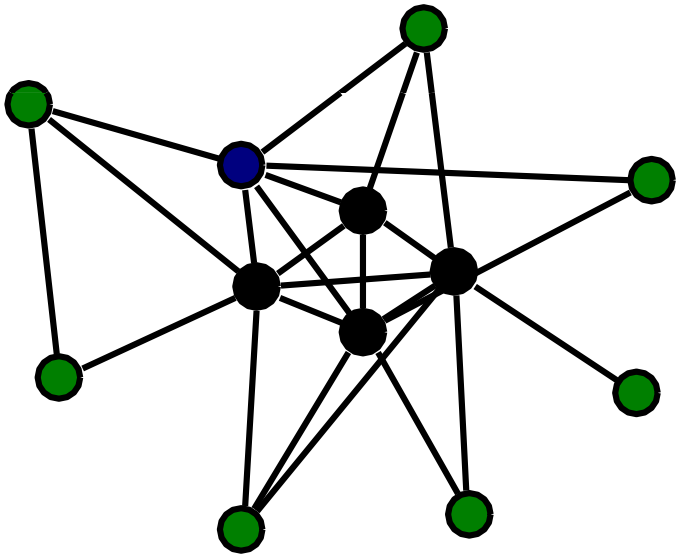
avg. dist. = 1.9



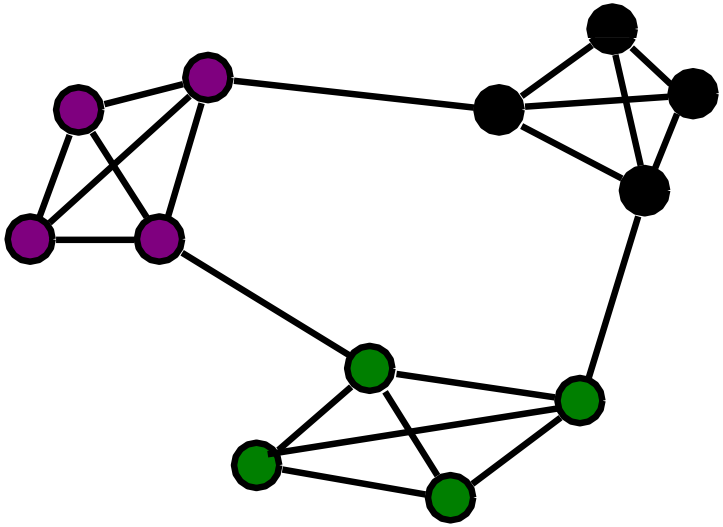
avg. dist. = 2.4

Diameter

- Maximum distance



Diameter = 3



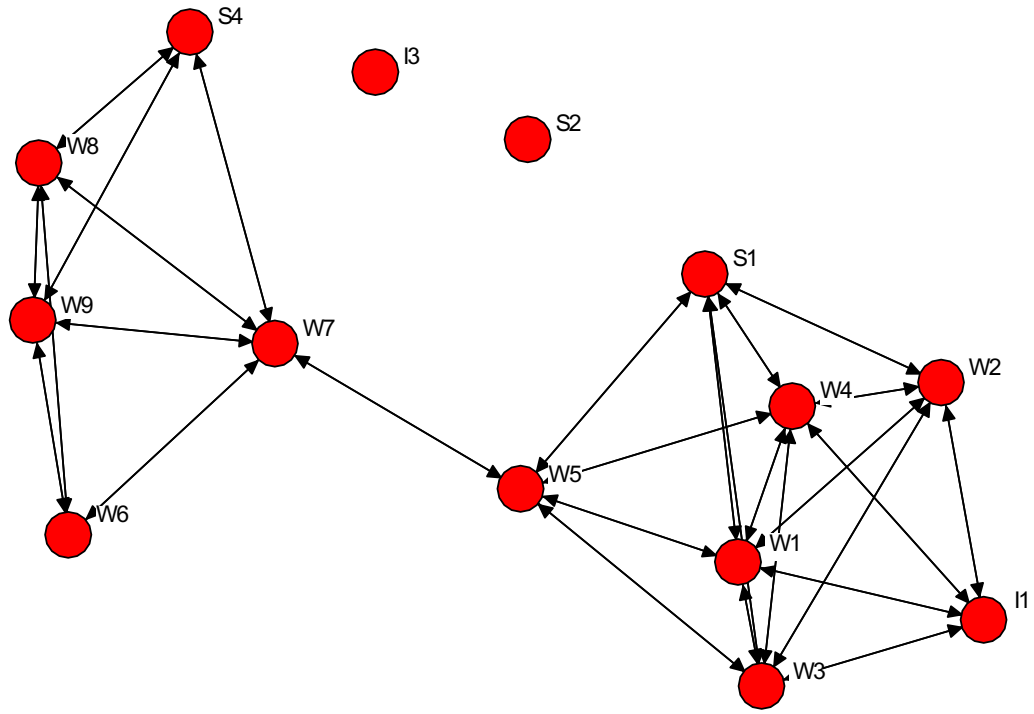
Diameter = 3

Fragmentation Measures

- Component ratio
- F measure of fragmentation
- Distance-weighted fragmentation ^{D}F

Component Ratio

- No. of components divided by number of nodes



Component ratio = 3/14 = 0.21

F Measure of Fragmentation

- Proportion of pairs of nodes that are unreachable from each other

$$F = 1 - \frac{\sum_{i \neq j} r_{ij}}{n(n-1)}$$

r_{ij} = 1 if node i can reach node j by a path of any length
r_{ij} = 0 otherwise

- If all nodes reachable from all others (i.e., one component), then F = 0
- If graph is all isolates, then F = 1

Computation Formula for F Measure

- No ties across components, and all reachable within components, hence can express in terms of size of components

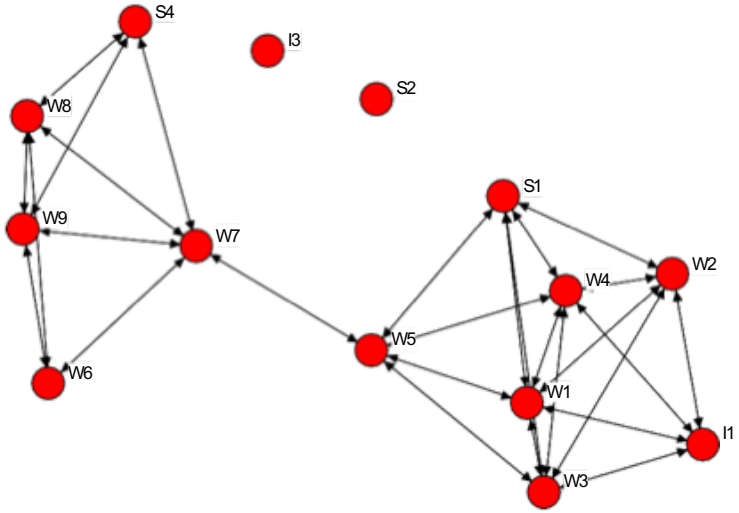
$$F = 1 - \frac{\sum_k s_k (s_k - 1)}{n(n-1)}$$

S_k = size of kth component

Computational Example

Games Data

Comp	Size	Sk(Sk-1)
1	1	0
2	1	0
3	12	132
	14	132
		0.2747 = 1 – (132/(14*13)) = F



Distance-Weighted Fragmentation

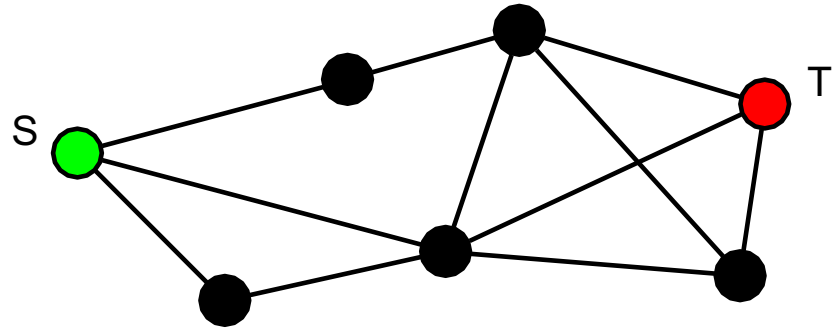
- Use the reciprocal of distance
  - letting $1/\infty = 0$

$$^{D}F = 1 - \frac{\sum_{i \neq j} \frac{1}{d_{ij}}}{n(n-1)}$$

- Bounds
  - lower bound of 0 when every pair is adjacent to every other (entire network is a clique)
  - upper bound of 1 when graph is all isolates

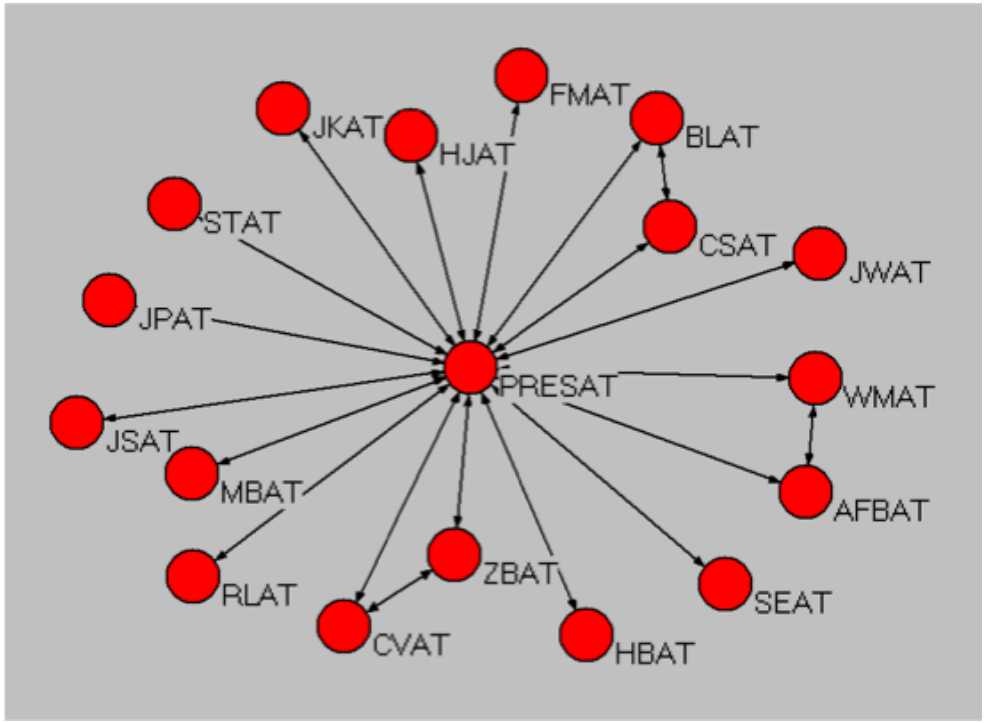
Connectivity

- Line connectivity λ is the minimum number of lines that must be removed to discon- nect network
- Node/point connectivity κ is minimum number of nodes that must be removed to discon- nect network



Centralization

- Degree to which network revolves around a single node

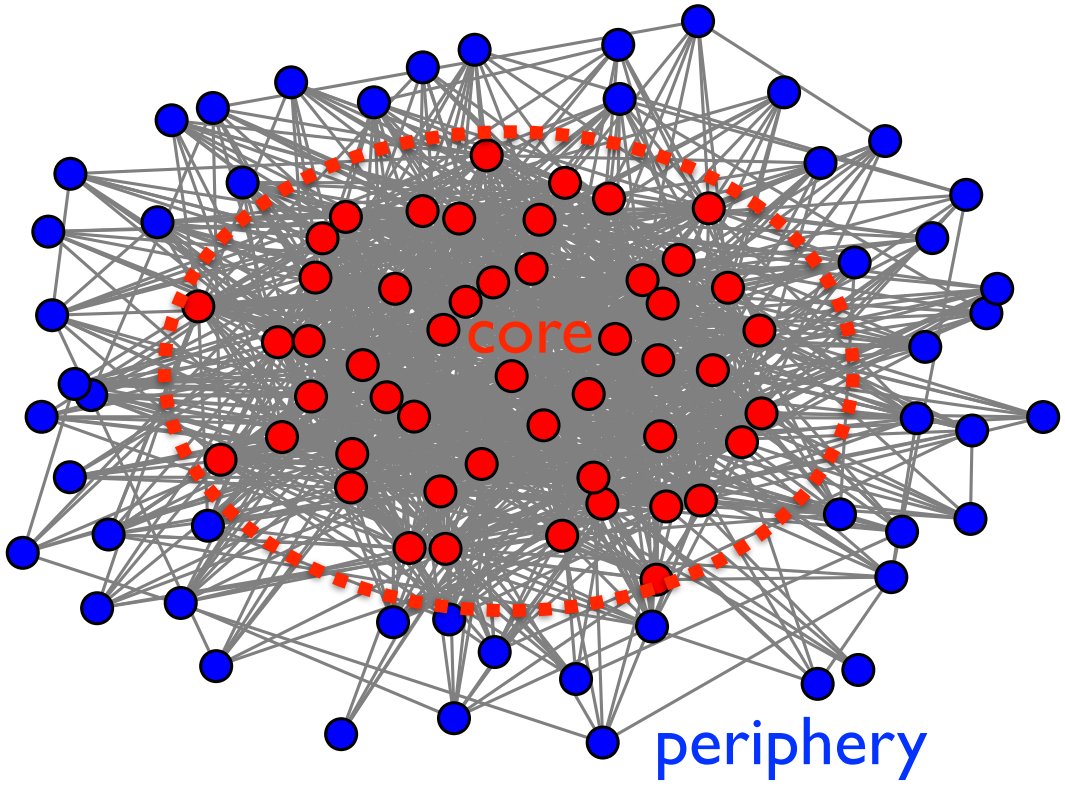


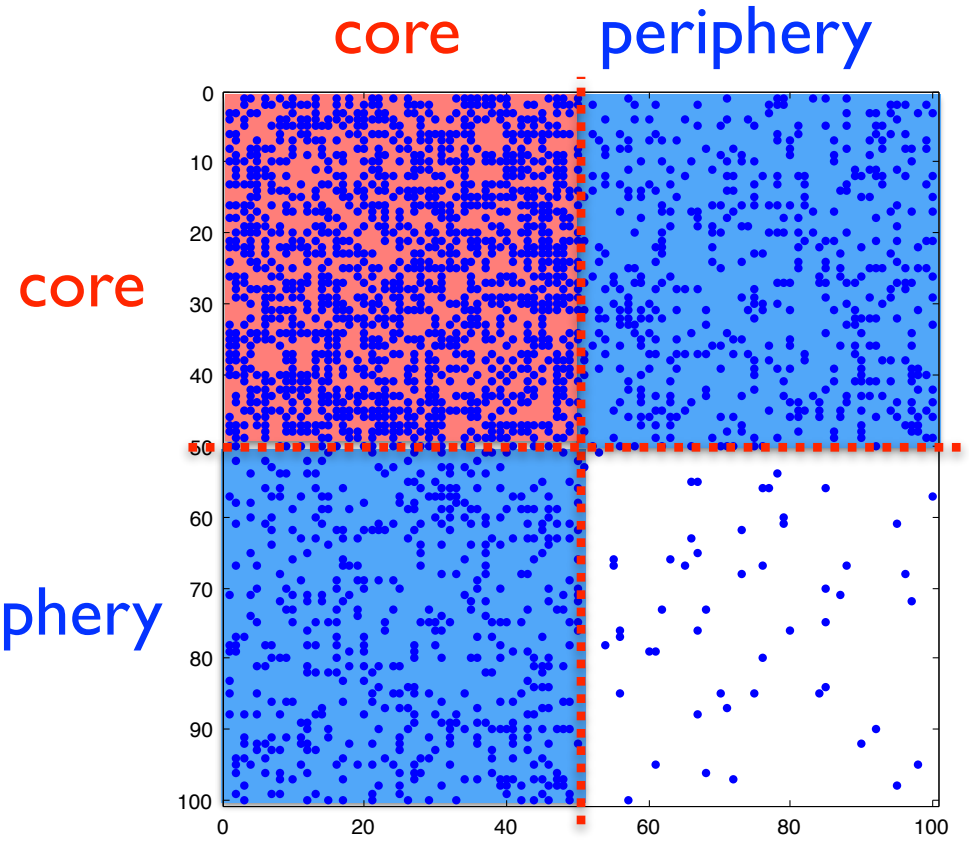
Carter admin.
Year 1

Core-Periphery Models

- A core periphery structure has a single cohesive subgroup with a set of other nodes, loosely connected to the core
- Core members interact with (lots of) other core members
- Periphery members interact with (a few) core members
- Periphery members rarely interact with each other

Finding Core/Periphery Structures



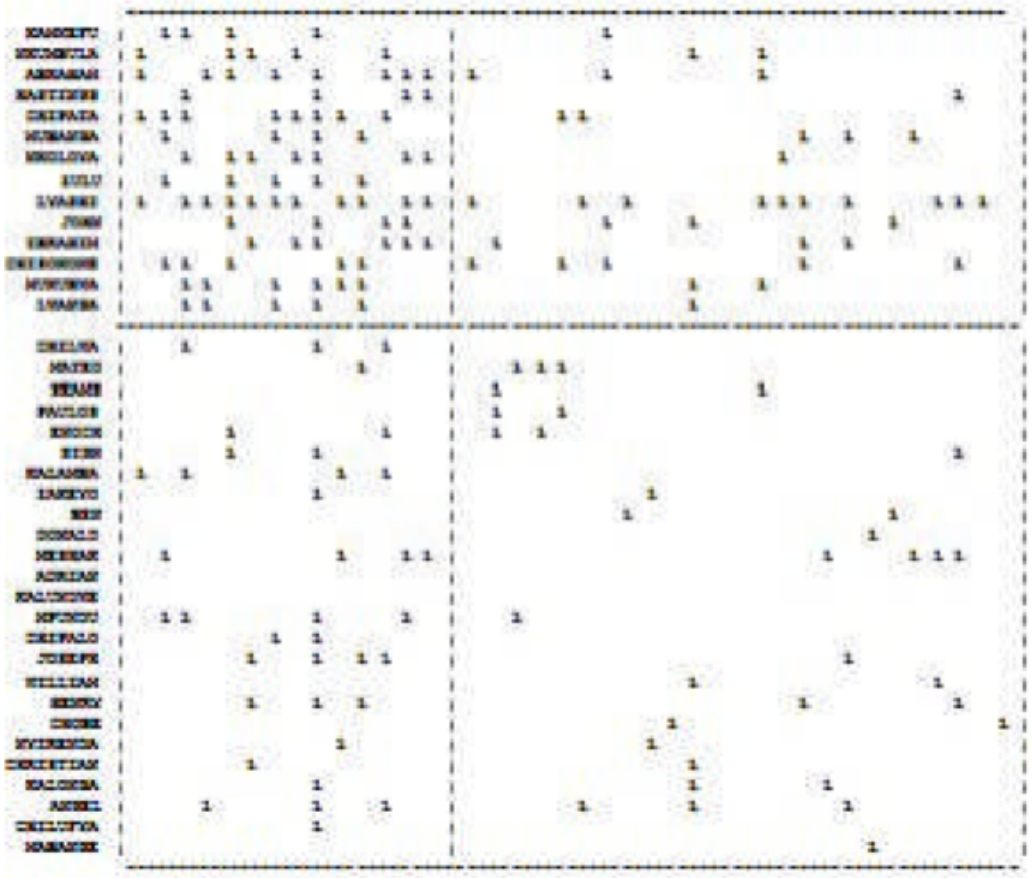


Core Periphery Block Model

Basic Idea:

- A *module* or *community* is a collection of nodes defined by how its *edges* behave:
  - **Edge Density:** For social networks, we expect edge density to be greater within a community than without. (Assortative Community)
  - **Edge Weight:** For coexpression networks, we expect the correlations to be higher within a functional module than without.
  - Etc.

Core Periphery Block Model



- Density Matrix

	1	2
1	0.451	0.106
2	0.106	0.057

Continuous Core/Periphery

- Calculate a “Coreness” vector C, in which c_i is the likelihood actor i is in the core
- Run a “Concentration” score to determine what is most appropriate core size
  - Basically correlate “coreness” values to ideal partition of core (1) and periphery (0)
    - Runs other measures as well
  - Pick the size with the highest correlation
- Create an “expected value” matrix which is CC^T (product of each dyad’s coreness)

Dyadic Cohesion

- Adjacency
  - Strength of tie
  - Reciprocity
- Reachability
  - A path exists or does not (usually as $1/d_{ij}$)
- Distance
  - Length of shortest path between two nodes
  - # Geodesics (how many paths of this length)
- Multiplexity
  - Number of ties of different relations linking two nodes
- Number of paths linking two nodes
  - Edge independent
  - Node independent

Average is density

1- f(Average) is fragmentation
Or distance weighted fragmentation

Average is average distance

Minimum is line connectivity

Minimum is point connectivity

Cohesive Subgroups & Communities

Broadly: “a group of nodes that are *relatively densely* connected to *each other* but sparsely connected to *other* dense groups in the network” Porter et al. 2009

No universal definition! But some ideas are:

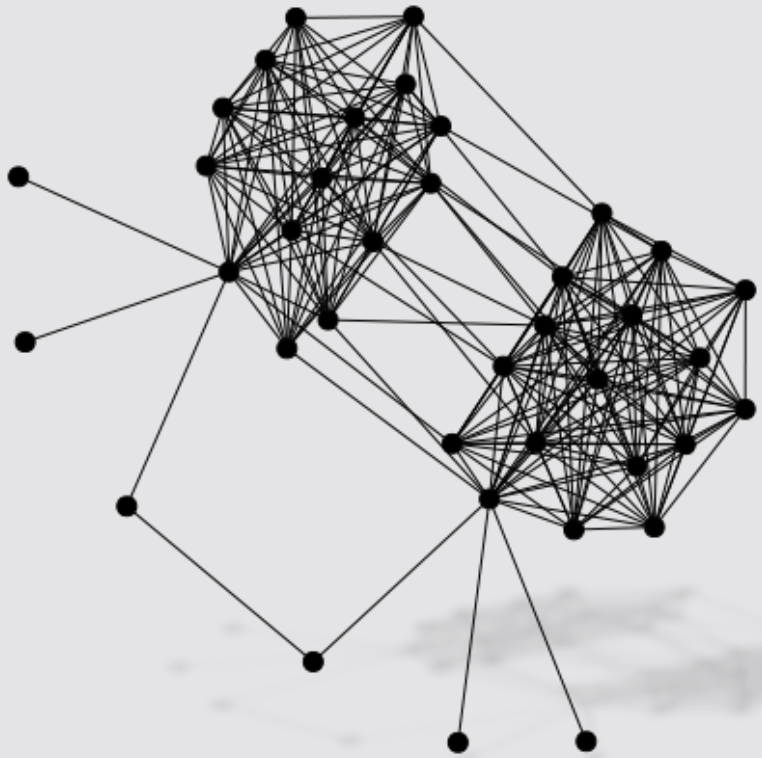
- A community should be densely connected
- A community should be well-separated from the rest of the network
- Members of a community should be more similar among themselves than with the rest

Most common..

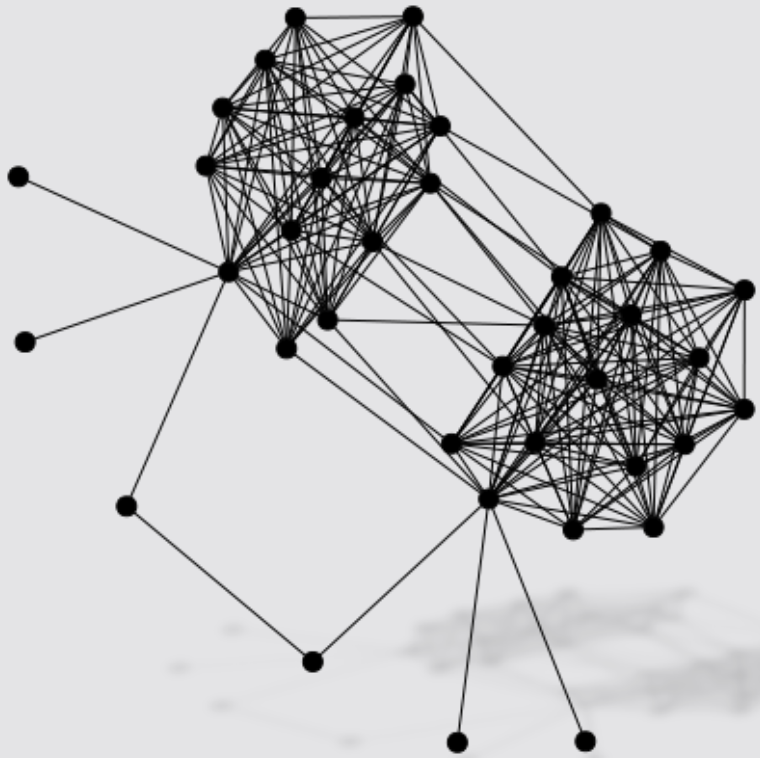
nr. of intra-cluster edges > nr. of inter-cluster edges

Typology of network communities

1. Cohesive subgroups
2. Similarity based clustering (agglomerative)
3. Graph partitioning (divisive)



Imagine this Graph ….

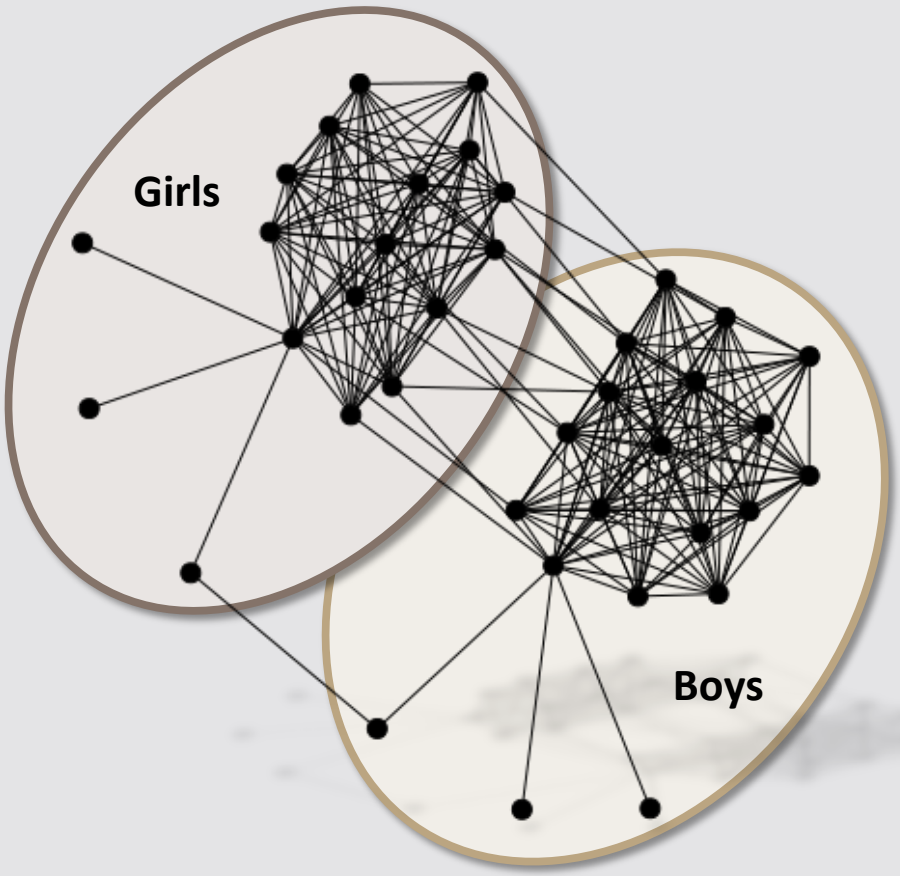


Vertices: People
Edges: Friendship

What factors might affect the formation of friendships in a high school social network?

Ideas: Age, Gender, Class, Race, Interests

How might we assign communities to this network?

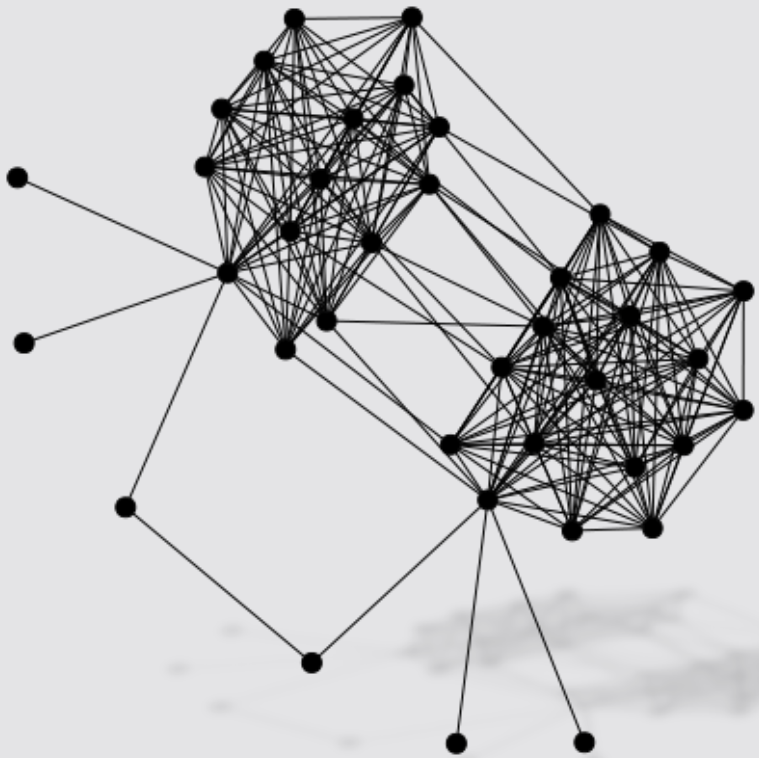


Vertices: People
Edges: Friendship

What factors might affect the formation of friendships in a high school social network?

Ideas: Age, Gender, Class, Race, Interests

How might we assign communities to this network?

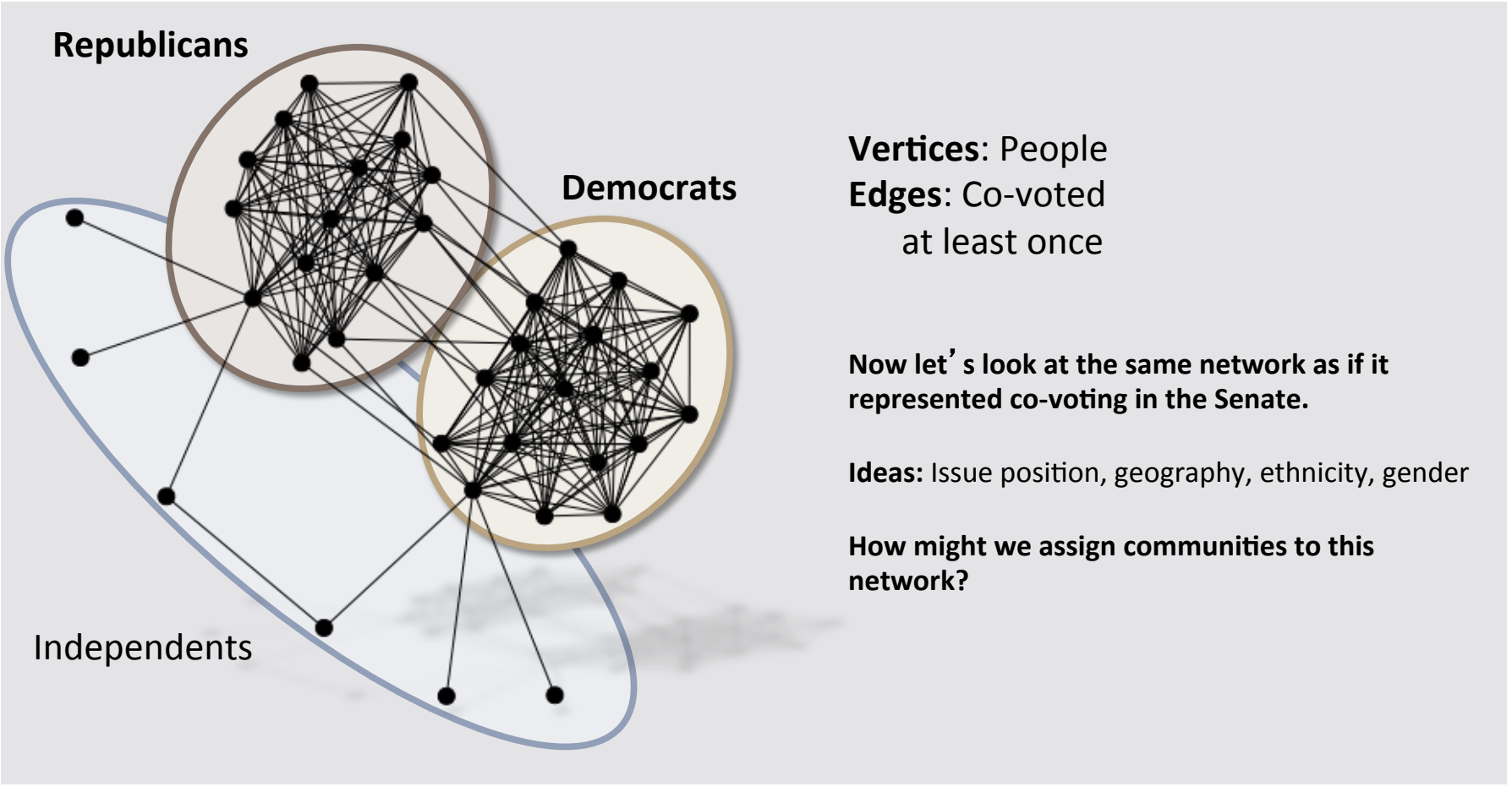


Vertices: People
Edges: Co-voted
at least once

Now let's look at the same network as if it represented co-voting in the Senate.

Ideas: Issue position, geography, ethnicity, gender

How might we assign communities to this network?

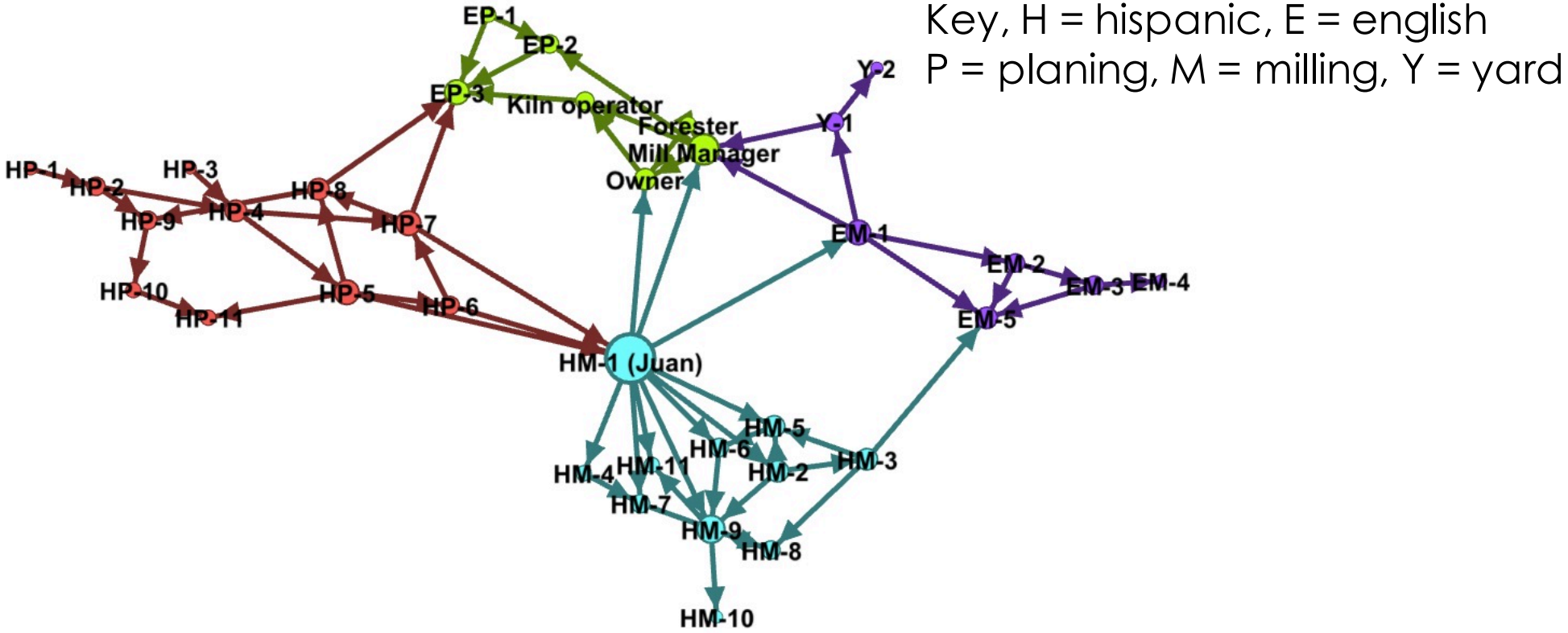


Vertices: People
Edges: Co-voted
at least once

Now let's look at the same network as if it represented co-voting in the Senate.

Ideas: Issue position, geography, ethnicity, gender

How might we assign communities to this network?

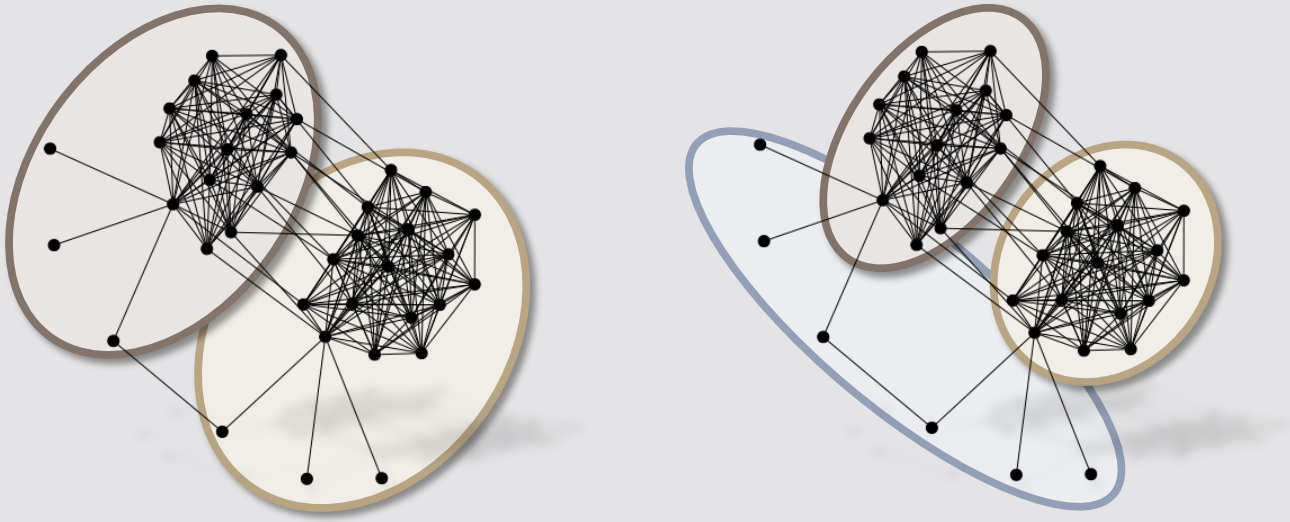


- The management at the sawmill was having difficulty persuading the workers to adopt a new plan, even though everyone would benefit. In particular the Hispanic workers (H) were reluctant to agree. The management called in a sociologist who mapped out who talked to whom regularly. Then they suggested that the management talk to Juan and have him talk to the Hispanic workers. It was a success, promptly everyone was on board with the new plan. Why?

context matters

Note that we have assigned community membership differently despite observing the *same graph!*

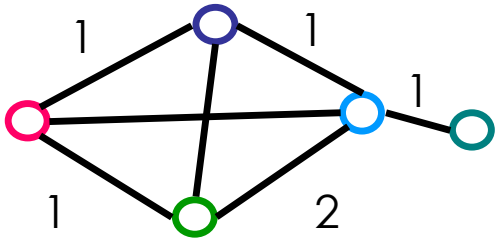
Community detection is not a concept that can be divorced from context.

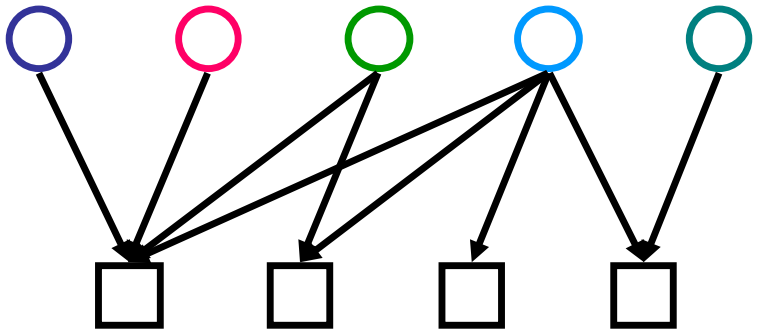


context matters – why do we observe communities at all?

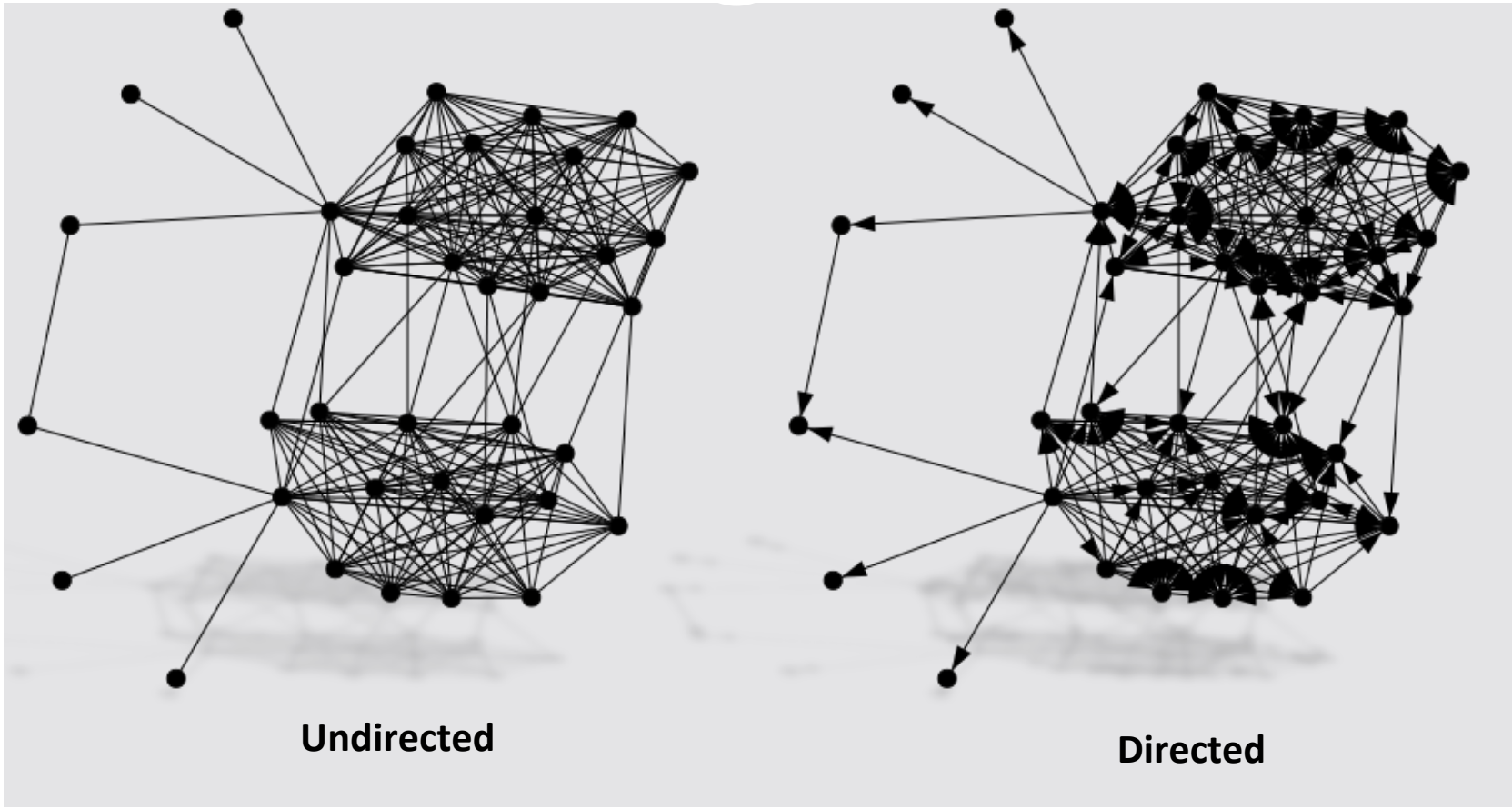
they arise out of an affiliation network! the one-mode projection we observe is an embedding of a multidimensional network that exists.

- otherwise known as
  - membership network
    - e.g. board of directors
  - hypernetwork or hypergraph
  - bipartite graphs
  - interlocks





practical aspects



Many methods:

- do not incorporate direction;
- allow for bidirected edges;
- may implement same method with or without support for directed egdes

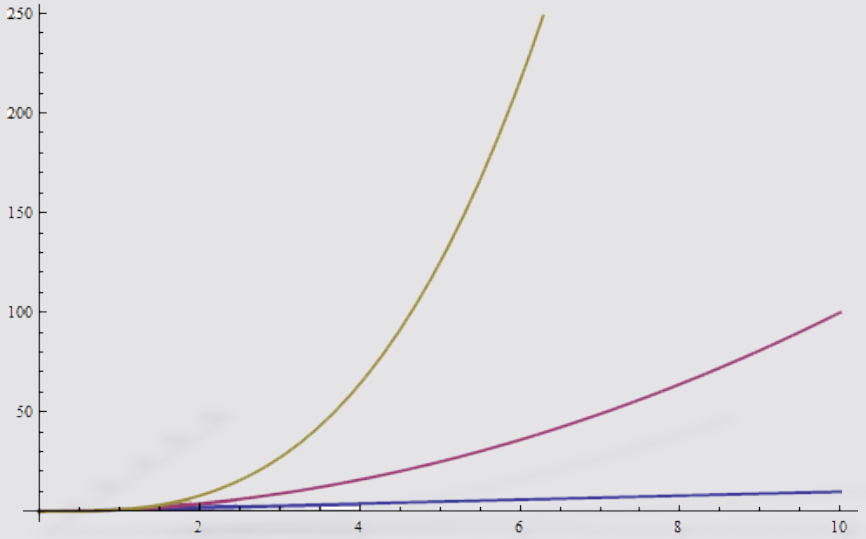
practical aspects

In computational complexity, "Big-O notation" conveys information about how time and storage costs scale with inputs.

- *O(1)*: constant - independent of input
- *O(n)*: scales linearly with the size of input
- *O(n^2)*: scales quadratically with the size of input
- *O(n^3)*: scales cubically with the size of input

These terms often occur with *log n* terms and are then given the prefix "quasi-."

For graph algorithms, the input *n* is typically
- $|V|$, the number of vertices
- $|E|$, the number of edges



Computation complexity mainly focused on two resources:

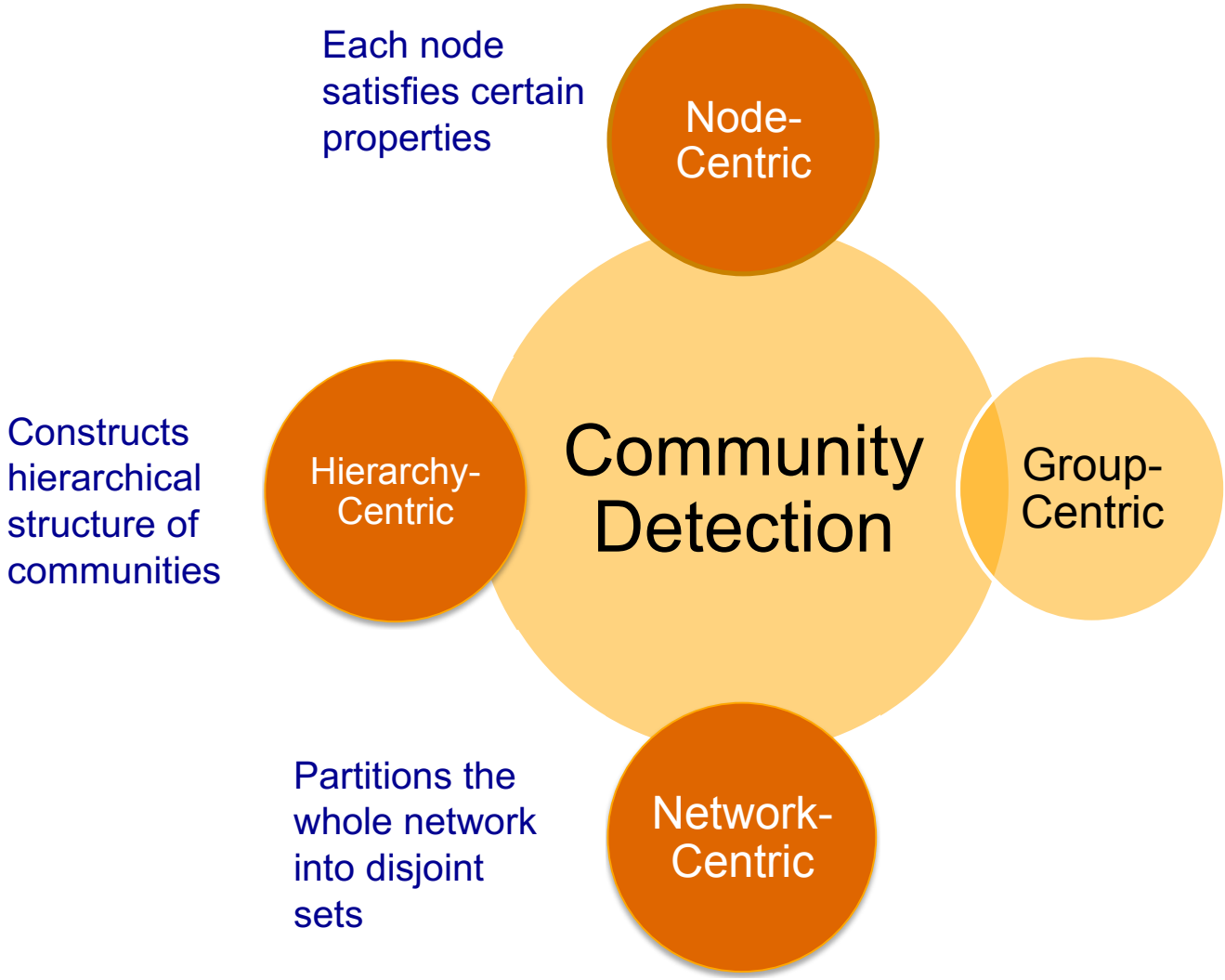
1. **Time** – how long does it take to perform sequence of operations?
2. **Storage** – how much space does it take to store our problem?

We tend to communicate both through "Big-O notation".

Cohesive Subgroups: A Typology

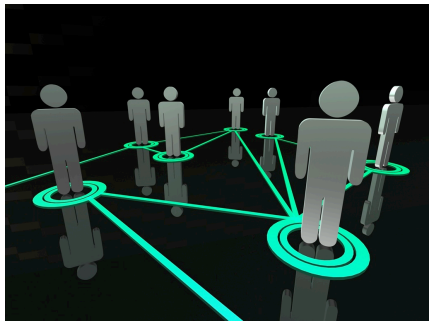
	Found by algorithm (**input** data driven)	Found by finding sets with **output** properties
Network / Graph theory	*Graph-theoretic data driven algorithms* Newman-Girvan	*Formal definitions of sociological groups* *{mathematical ethnography}* Clique, n-clique, n-clan, n-club, k-plex, ls-set, lambda-set, k-core, component
Proximities / Clustering	*Multivariate clustering analysis methods* Johnson's Hierarchical clustering; k-means; MDS	*Formal definitions of abstract clusters* Combinatorial optimization Factions (Core-Periphery)

taxonomy of communities



Basics of communities

We focus on the mesoscopic scale of the network



Microscopic Mesoscopic Macroscopic

Fundamental Hypotheses of communities

H1: A network's community structure is uniquely encoded in its wiring diagram

H2: **Connectedness Hypothesis** – a community corresponds to a connected subgraph

H3: **Density Hypothesis** – communities correspond to locally dense neighbourhoods of a network;

H4: **Random Hypotheses**: randomly wired networks are not expected to have a community structure;

H5: **Maximal Modularity Hypotheses**: the partition with the maximum modularity *M* for a given network offers the optimal community structure

Fundamental Hypotheses of communities

Strong and weak communities

Consider a connected subgraph *C* of N_c nodes

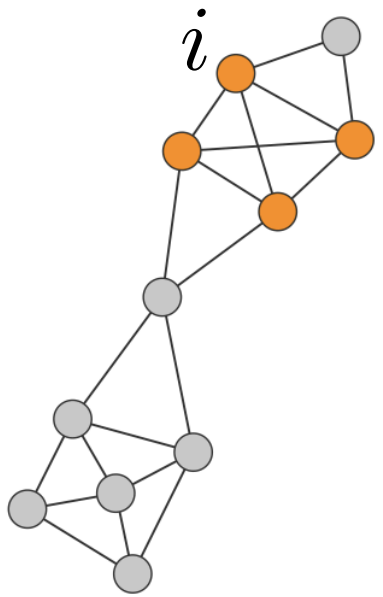
Internal degree, k_i^{int} *:* set of links of node *i* that connects to other nodes of the same community *C*.

External degree k_i^{ext}*:* the set of links of node *i* that connects to the rest of the network.

If $k_i^{ext}=0$*:* all neighbors of *i* belong to C, and *C* is a good community for *i*.

If $k_i^{int}=0$, all neighbors of *i* belong to other communities, then *i* should be assigned to a different community.

$$k_i^{int} = 3$$
$$k_i^{ext} = 1$$



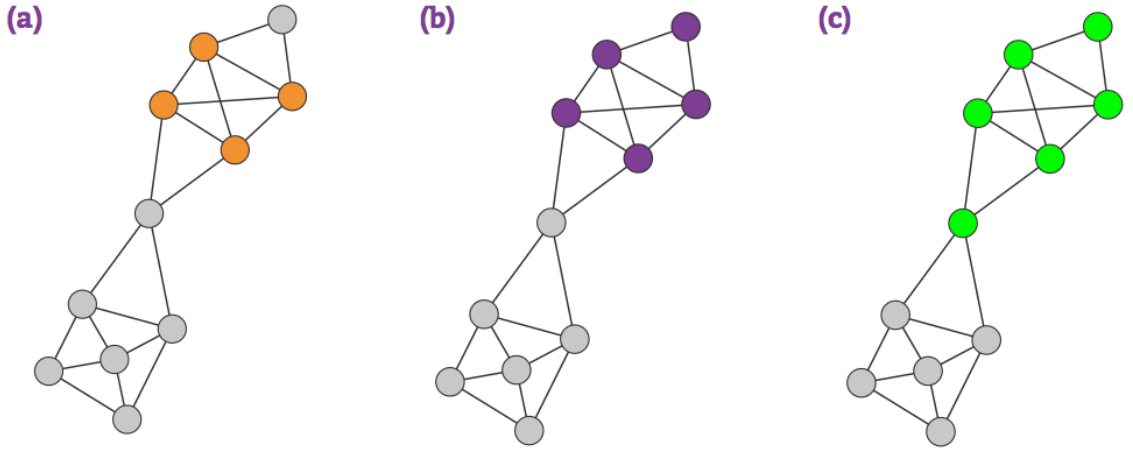
Fundamental Hypotheses of communities

Strong community:
Each node of *C* has more links within the community than with the rest of the graph.

$$k_i^{\text{int}}(C) > k_{\text{i}}^{\text{ext}}(C)$$

Weak community:
The total internal degree of *C* exceeds its total external degree,

$$\sum_{\text{i}\in\text{C}} k_i^{\text{in}}(C) > \sum_{\text{i}\in\text{C}} k_{\text{i}}^{\text{out}}(C)$$



Clique *Strong* *Weak*

Node-Centric | Community Detection (Cohesive subgroups)

Node-Centric | Community Detection

Defined by graph-theoretic characteristics of resultant sets, where nodes must satisfy different properties:

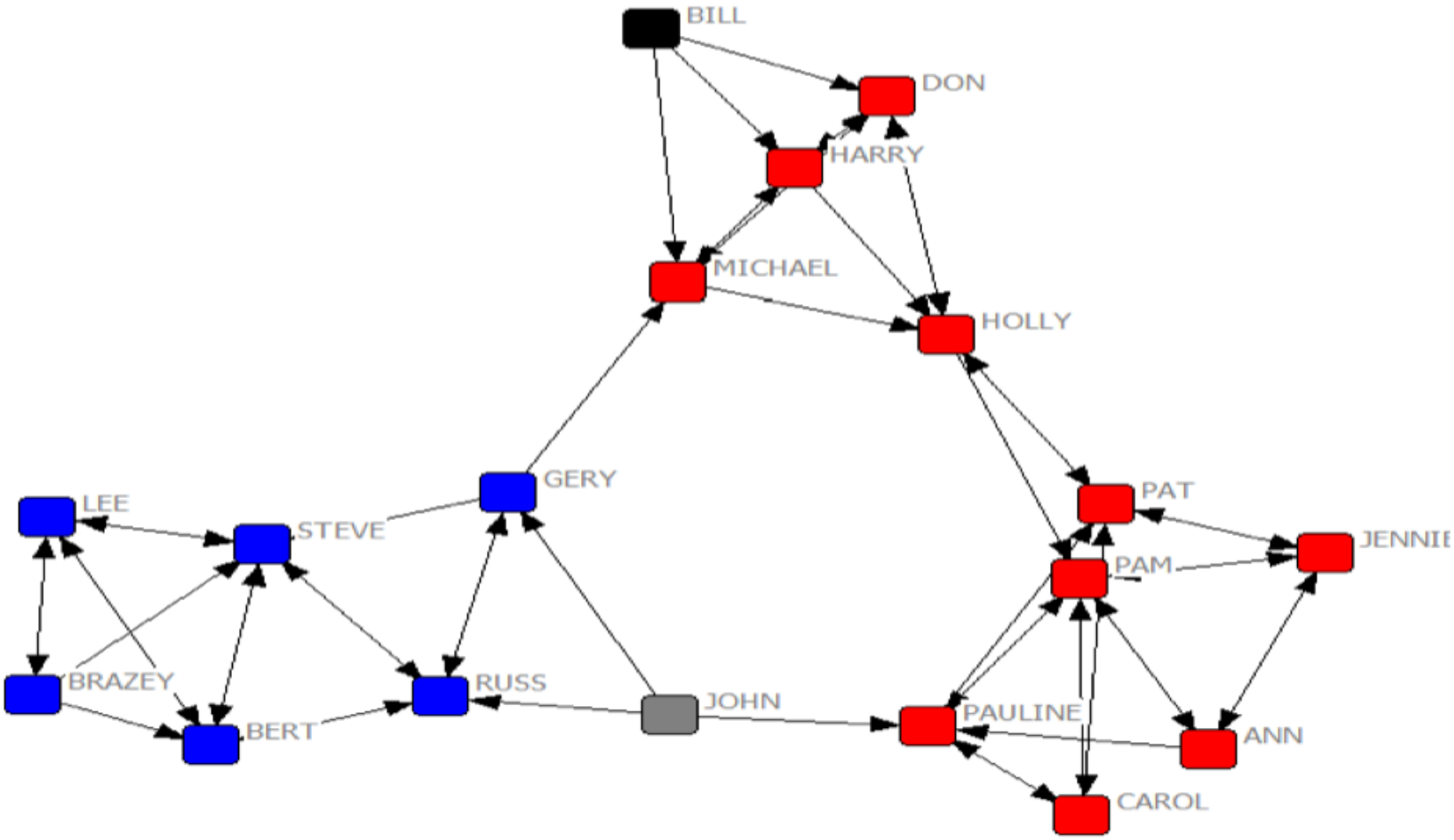
- **Complete Mutuality** [everybody in the group knows everybody else]
  - components
  - cliques
- **Reachability of members** [individuals are separated by at most n hops]
  - n-clique, n-clan, n-club
- **Nodal degrees** [everybody in the group has links to at least *k* others in the group]
  - k-plex, k-core
- **Relative frequency of *within-outside* ties** [subgroup members v non-members]
  - LS sets, Lambda sets

complete mutuality | **components**

- Maximally **connected** subgraph
  - In undirected graphs, it just means everyone's connected to everyone else
  - In digraphs there are strong and weak components:
    - **Strong components** mean everyone can reach everyone else, even when considering the one-way streets in the network
    - **Weak components** means, if we ignore the directionality of the ties, everyone is reachable by everyone else

Campnet

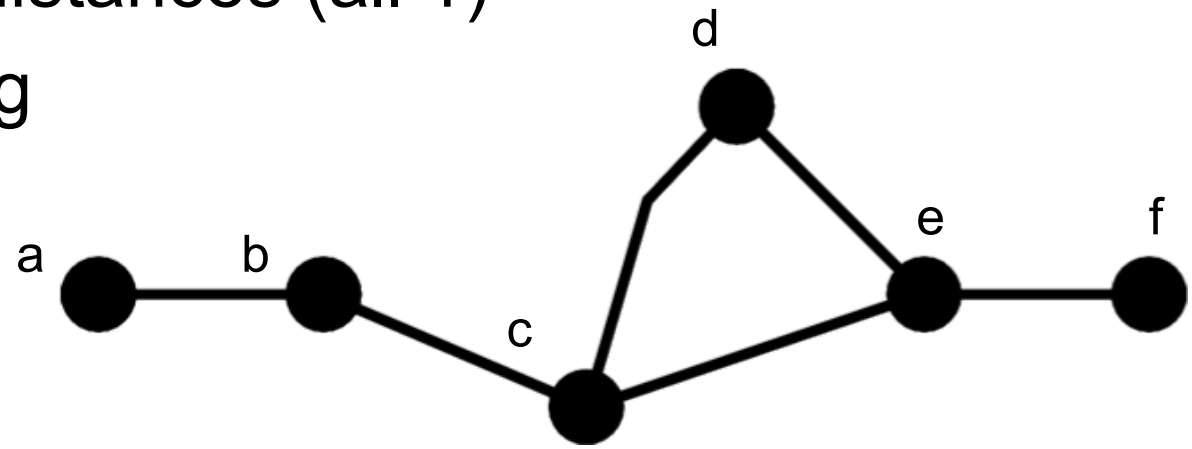
Colored by Strong Components



complete mutuality | **cliques**

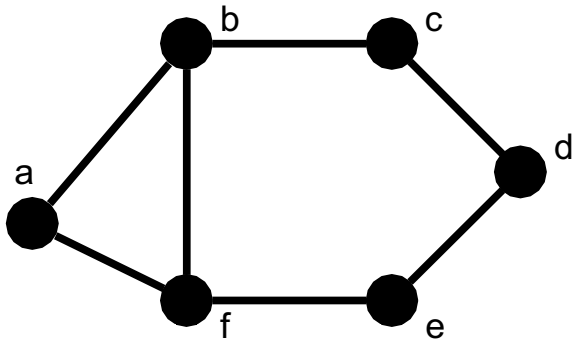
- Definition
  - Maximal, complete subgraph
  - Set S s.t. for all u,v in S, (u,v) in E
- Properties
  - Maximum density (1.0)
  - Minimum distances (all 1)
  - overlapping
  - Strict

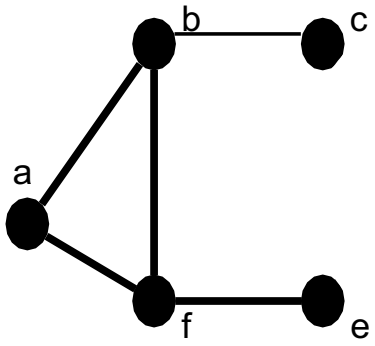
{c,d,e} is the only clique



Subgraphs

- Set of nodes
  - Is just a set of nodes
- A subgraph
  - Is set of nodes together with ties among them
- An induced subgraph
  - Subgraph defined by a set of nodes
  - Like pulling the nodes and ties out of the original graph

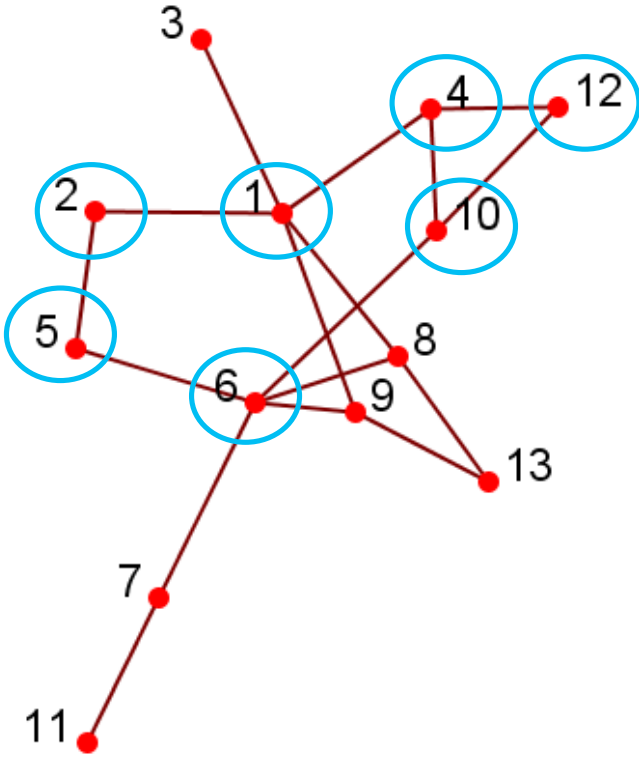




Subgraph induced by {a,b,c,f,e}

Geodesic

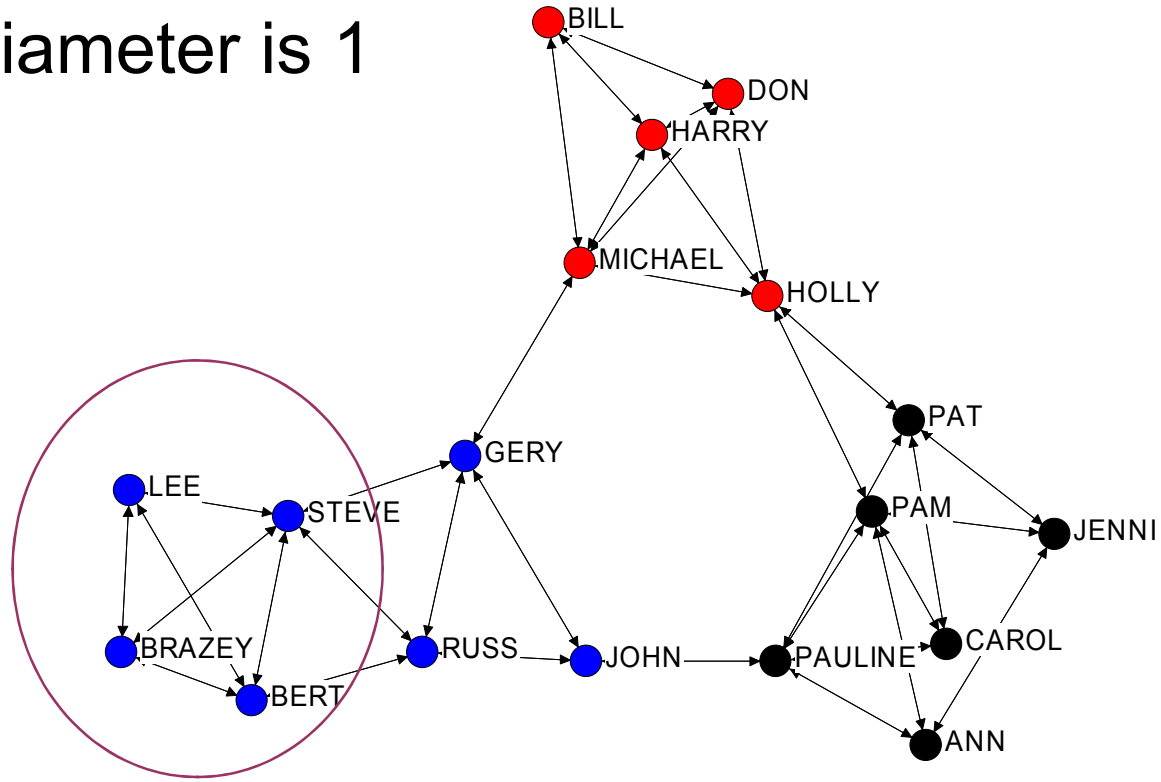
- Reachability is calibrated by the Geodesic distance
- Geodesic: a shortest path between two nodes (12 and 6)
  - Two paths: 12-4-1-2-5-6, 12-10-6
  - 12-10-6 is a geodesic
- Geodesic distance: #hops in geodesic between two nodes
  - e.g., d(12, 6) = 2, d(3, 11)=5
- Diameter: the maximal geodesic distance for any 2 nodes in a network
  - #hops of the longest shortest path



Diameter = 5

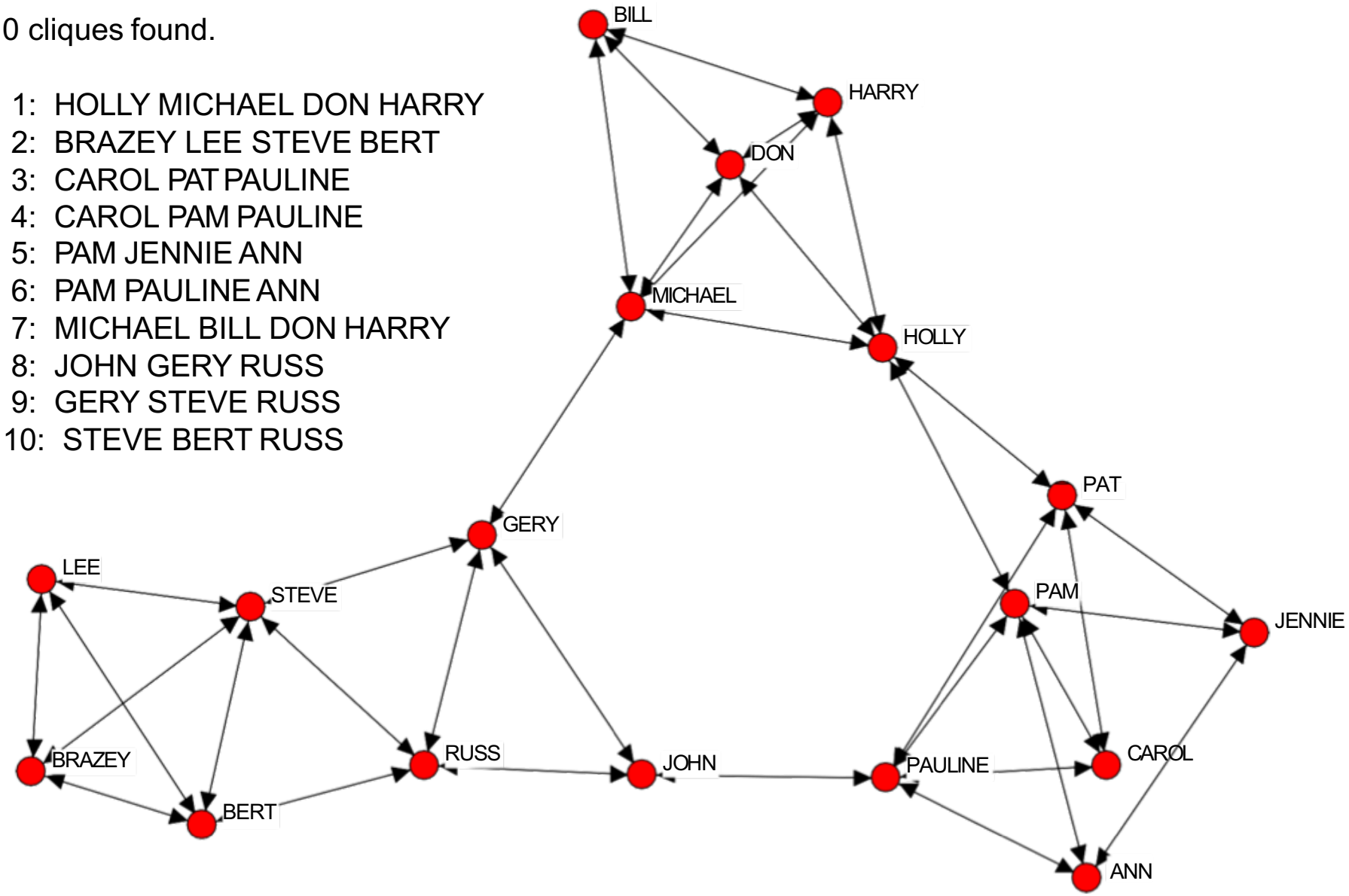
complete mutuality | clique

- A maximal **complete** subgraph
  - Everyone is adjacent to everyone else
  - Distance & Diameter is 1
  - Density is 1
- Limitations
  - Undirected
  - Binary
  - 3+ nodes



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



Problems with Cliques

- Very strict
- Not robust: one missing link can disqualify a clique
- Sometimes too many and overlapping;
- Not interesting
  - everybody is connected to everybody else
  - no core-periphery structure
  - no centrality measures apply
- Sometimes too few
  - This has lead to many kinds of relaxations.The distinctions between them are subtle, and not generally of **practical** importance.
    - We'll go through them, but don't worry about the nuances, just know multiple variants exist

Types of Relaxations

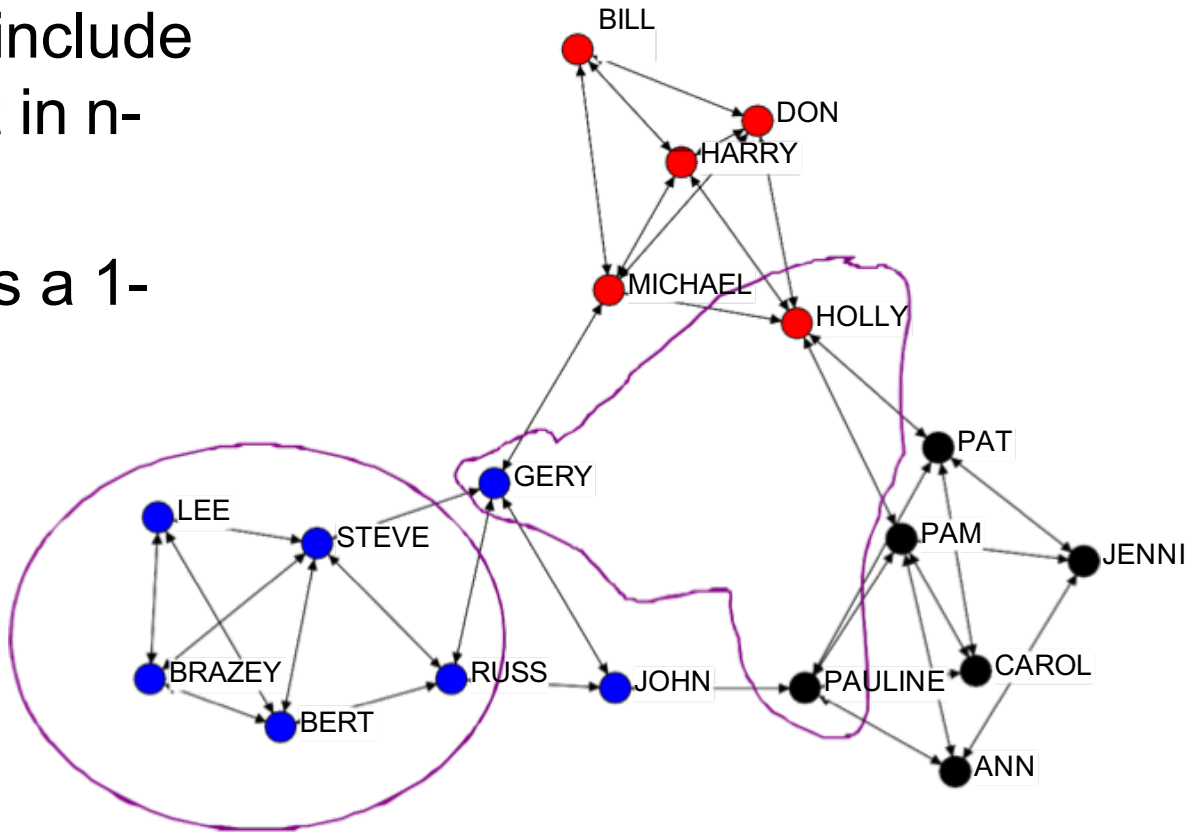
- Distance Relaxations (length of paths)
  - n-clique
  - n-clan
  - n-club
- Density Relaxations (number of ties)
  - k-plex
  - k-core

reachability of members | **n-clique**

- n-Clique
  - Maximal subset with all nodes within n steps of each other
    - Path can include nodes not in n-Clique
    - A Clique is a 1-Clique

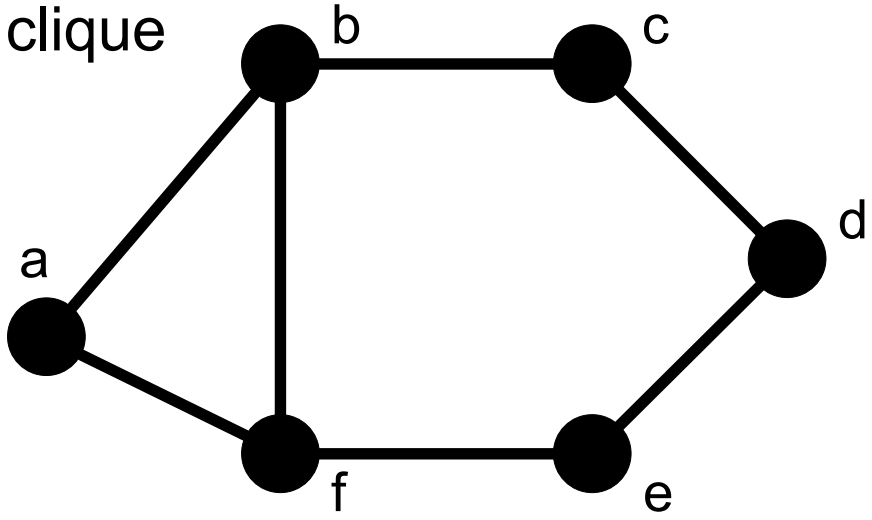
Is this a 2-Clique?
NO!
What about now?

But so is this!!!



reachability of members | n-clique

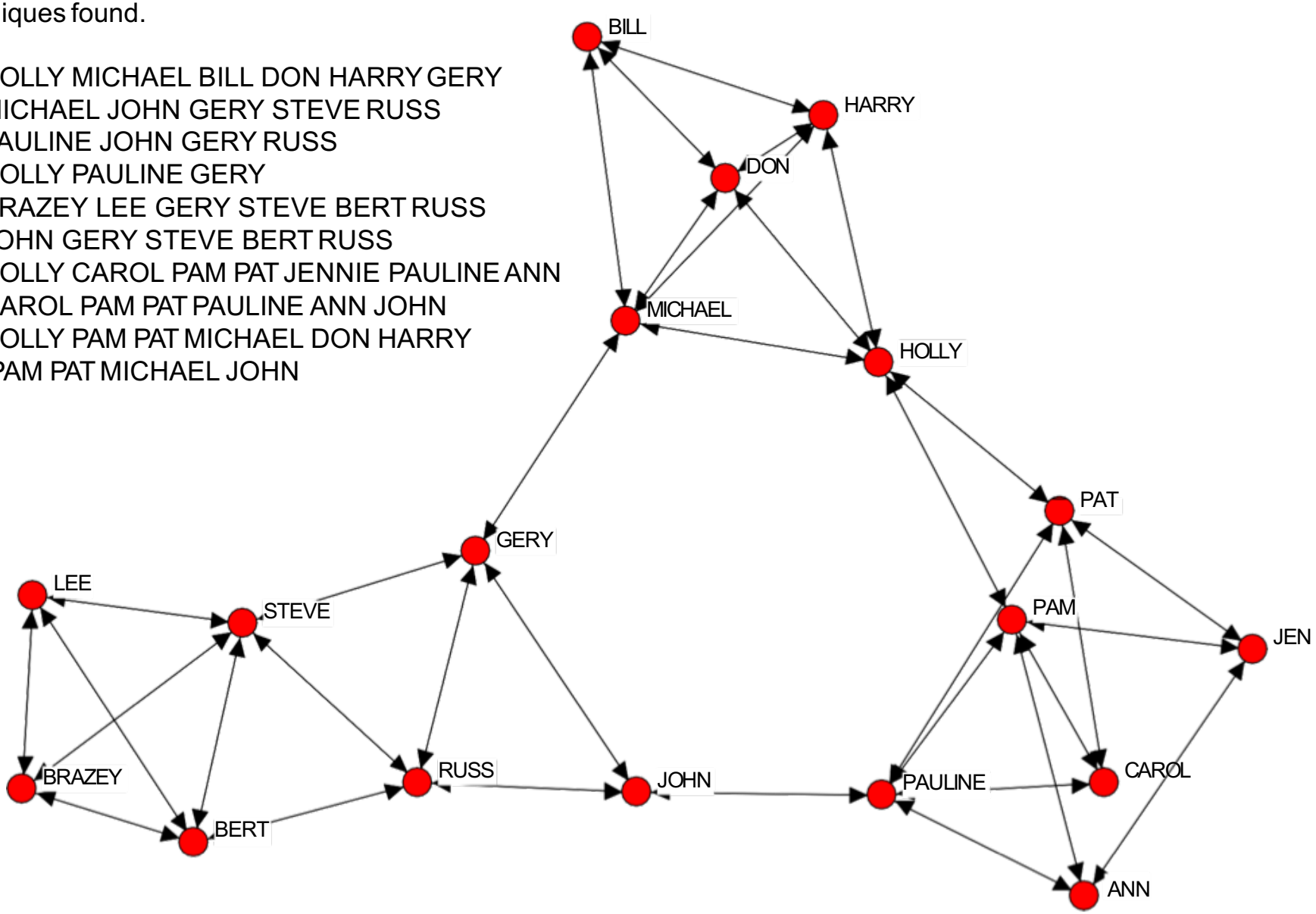
- Definition
  - Maximal subset s.t. for all *u,v* in S, d(*u,v*) <= n
  - Distance among members less than specified maximum
  - When n = 1, we have a clique
- Properties
  - Relaxes notion of clique
    - Avg distance can be greater than 1



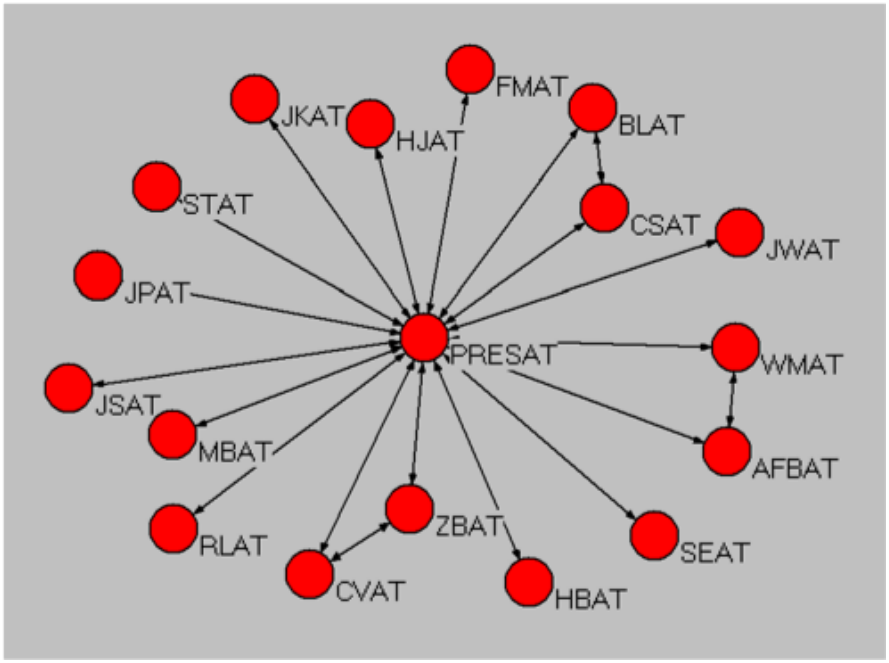
Is {a,b,c,f,e} a 2-clique?
yes

10 2-cliques found.

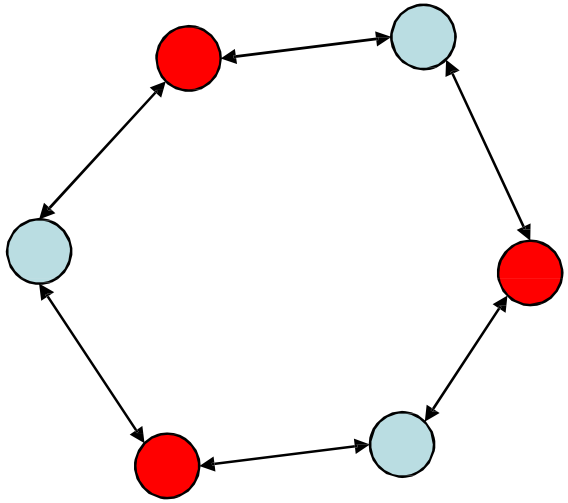
1: HOLLY MICHAEL BILL DON HARRY GERY
2: MICHAEL JOHN GERY STEVE RUSS
3: PAULINE JOHN GERY RUSS
4: HOLLY PAULINE GERY
5: BRAZEY LEE GERY STEVE BERT RUSS
6: JOHN GERY STEVE BERT RUSS
7: HOLLY CAROL PAM PAT JENNIE PAULINE ANN
8: CAROL PAM PAT PAULINE ANN JOHN
9: HOLLY PAM PAT MICHAEL DON HARRY
10: PAM PAT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



Some are counter-intuitive (And not necessarily cohesive)



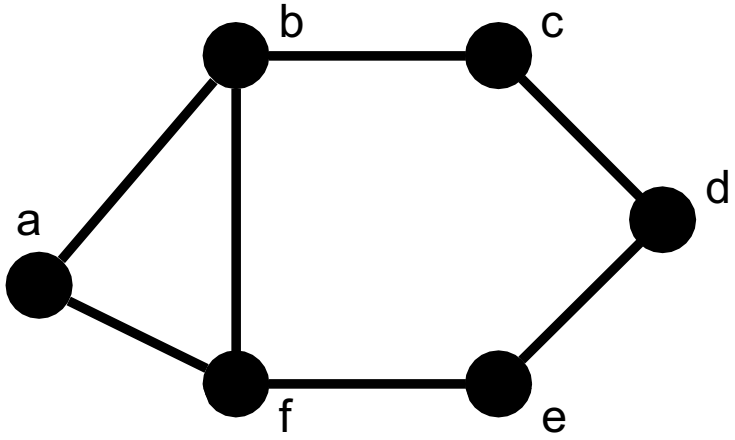
This is a 2-Clique

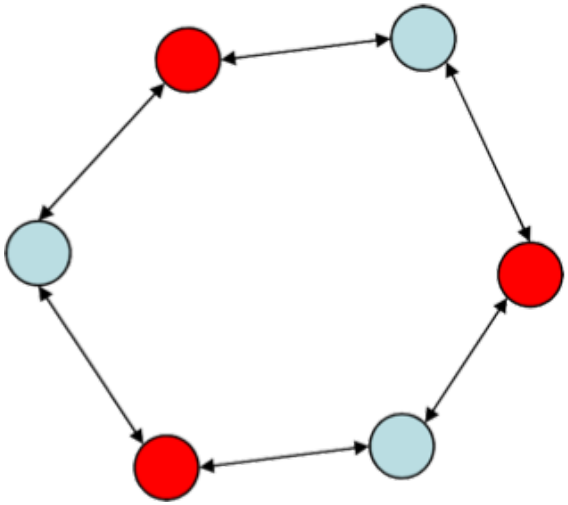


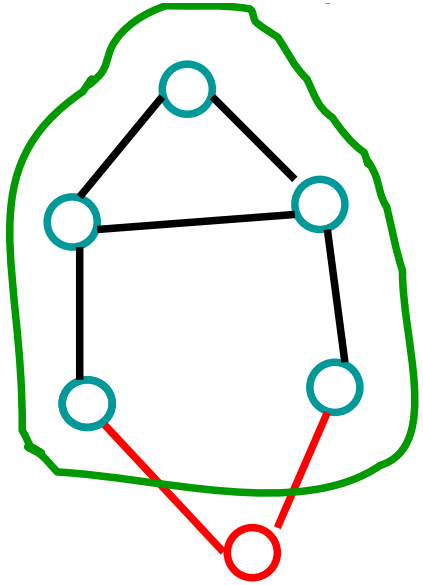
Red Nodes form a 2-Clique, so do Blues

Issues with N-Cliques

- Overlapping
  - {a,b,c,f,e} and {b,c,d,f,e} are both 2-cliques
- Membership criterion satisfiable through non-members
- Diameter may be greater than n
- n-clique may be disconnected (paths go through nodes not in subgroup)
- Even 2-cliques can be fairly non-cohesive
  - Both sets of alternating nodes belong to a different 2-clique but none are adjacent





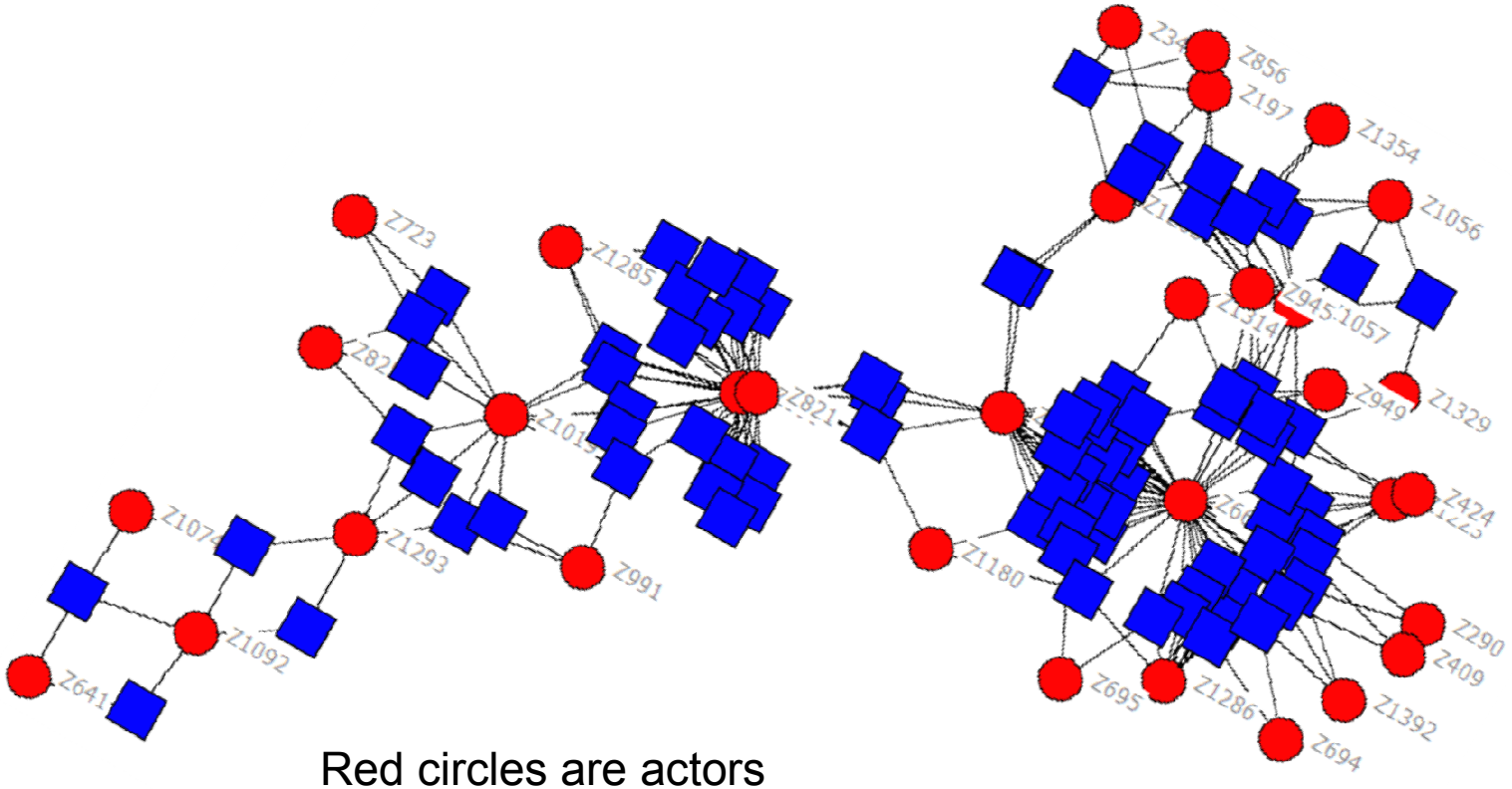


2 – clique
diameter = 3

path outside the 2-clique

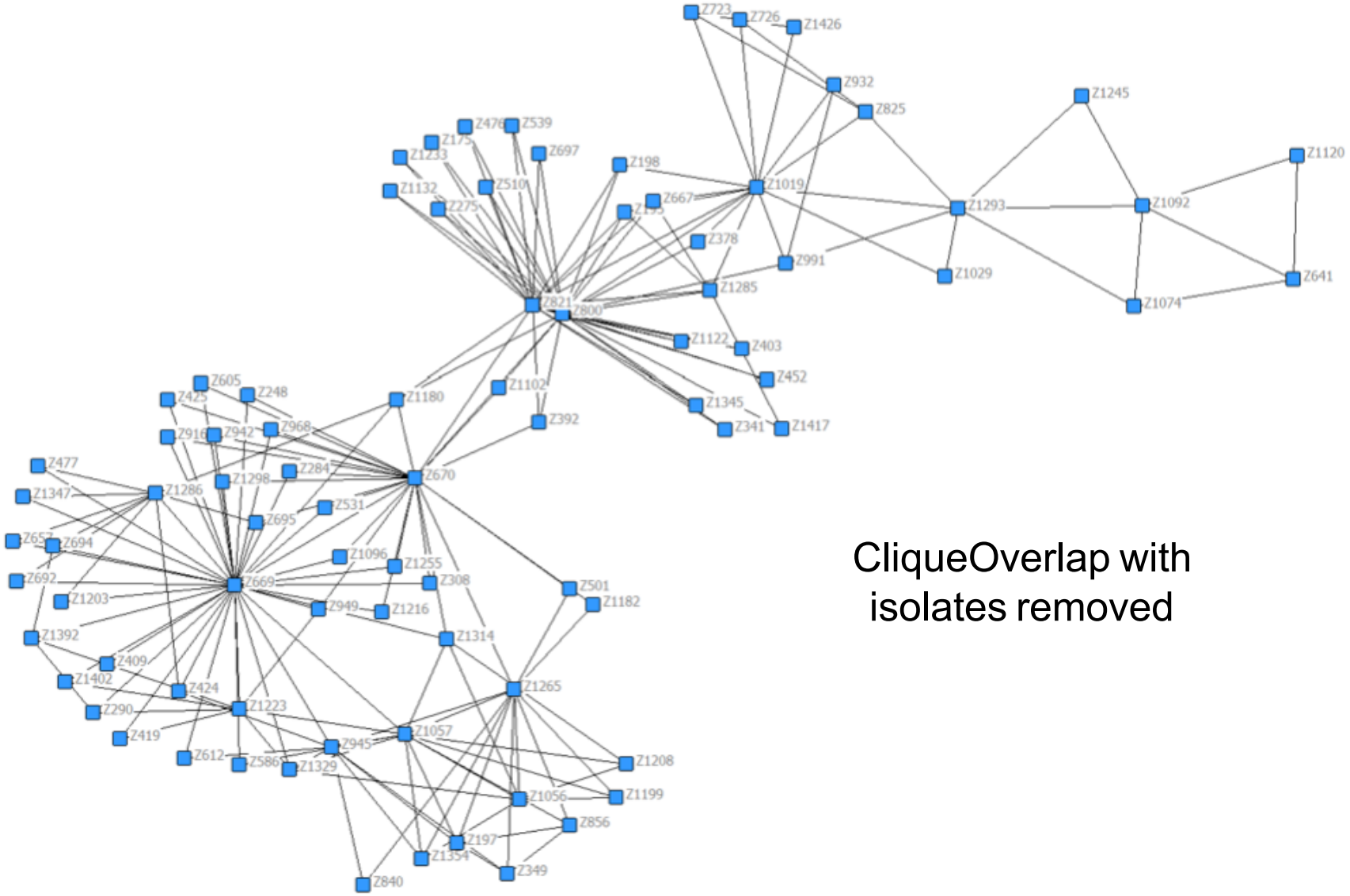
Many of these are (too) plentiful

- One way to process the information is to look at CliqueSets as a two-mode network



Red circles are actors
Blue squares are cliques

Or, Look at CliqueOverlap



CliqueOverlap with isolates removed

Loosen the density restriction

- n-Cliques (and the attempts to fix them, n-Clans, and n-Clubs) all start from the definition of Cliques and relax the distance requirement (all distances = 1) in varying ways:
  - **e.g. n-club: maximal subgraph of diameter 2**

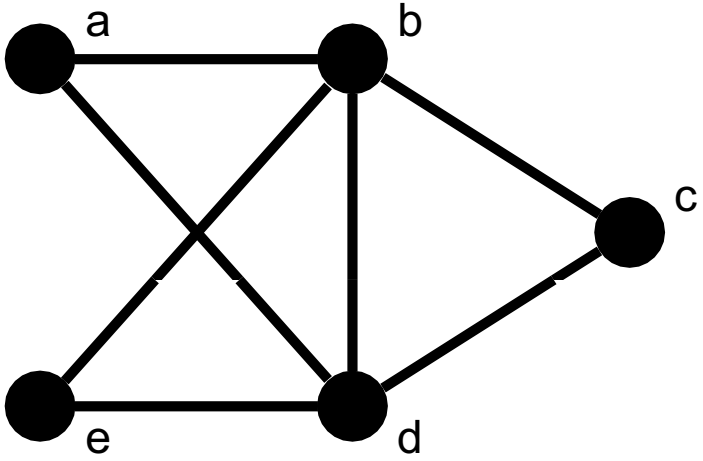
- But, Cliques also have **maximum density** (d = 1), and we can relax that definition instead.

- But for this, we must define the alpha operator,
  α, such that α(u,G) is the number of edges from node u to nodes in graph G

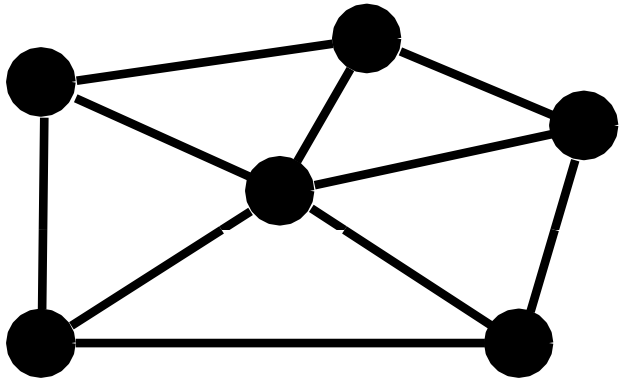
nodal degrees | **k-plex**

- k-Plex
  - A clique where members don't have to be connected to everyone else, just all but *k* members, or…
  - a [maximal] subgraph S s.t. for all *u* in S, $\alpha(u,S)$ >= |S|-k, where |S| is size of set S
    - All subsets of k-plexes are k-plexes (if non-maximal)
    - Get distance for free based on S, k.
      - If k < (|S|+2)/2 then diameter <= 2
    - Numerous & Overlapping
    - May be more intuitive than distance-based measures
    - A Clique is a 1-plex (We assume it not tied to itself)

K-Plex



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?

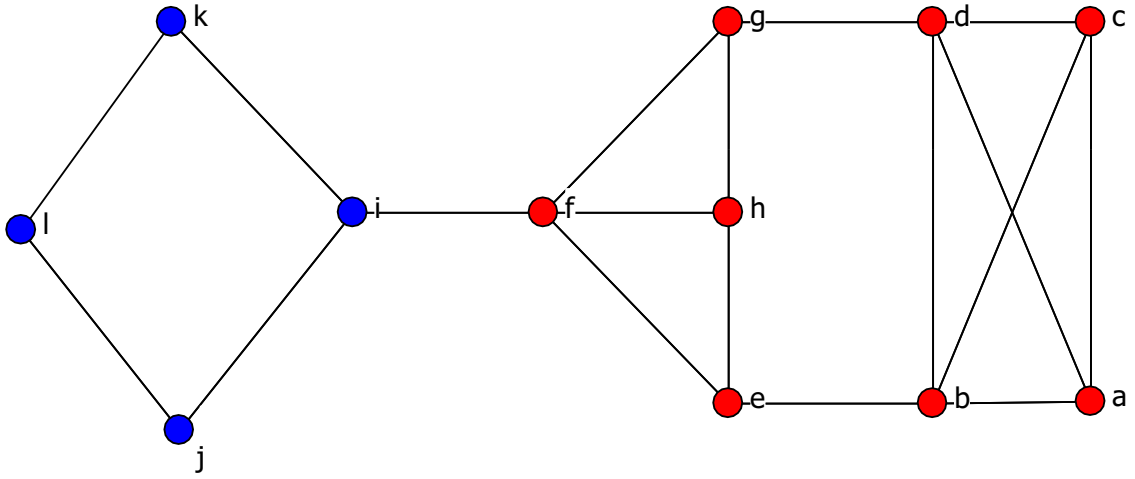
nodal degrees | **k-core**

- Sort of opposite approach from k-plex
  - Because the size of the group is not taken into account, k-cores are more directly about specifying how many ties MUST be present independent of how many nodes are in the core, whereas the k-plex is about how many may be missing.

- A k-Core is maximal subgraph within which all nodes have ties to at least k other nodes
  - All nodes in a components are at least 1-Cores
  - Each nodes is assigned a “core” which is the largest k-core to which it belongs (and it therefore also belongs to all lower cores that exist)
  - K-cores are hierarchical and form a partition
  - However, they may be disconnected

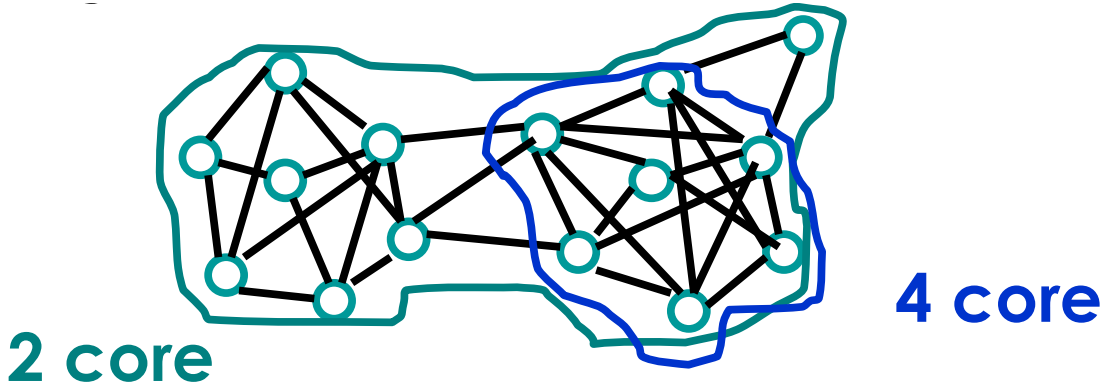
formal definition

- A k-core is a maximal subgraph such that for all u in S, α(u,S) >= k



  - All nodes are 2-core (and 1-core) Red nodes are 3-core.
- Great for analyzing large networks

but still too stringent…



node on top right only has 2 edges, so it is excluded from the 4 core group identified; the next k-core partition it can join is one that captures the whole network…

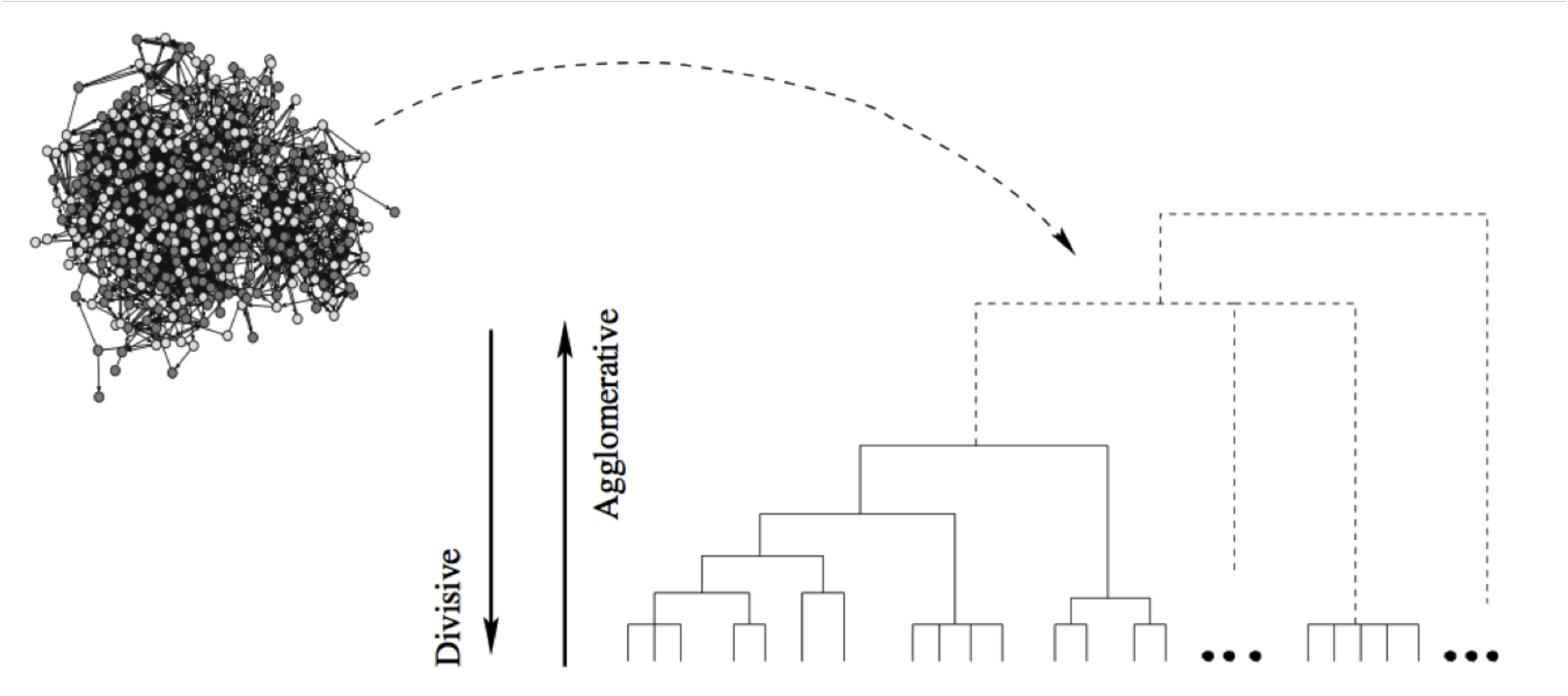
recap node-centric communities (cohesive subgroups)

- Each node has to satisfy certain properties
  - Complete mutuality
  - Reachability
  - Nodal degrees
  - Within-Outside Ties

- Limitations:
  - Too strict, but can be used as the core of a community
  - Not scalable, commonly used in network analysis with small-size network
  - Sometimes not consistent with property of large-scale networks
    - e.g., nodal degrees for scale-free networks

Network-Centric | [Agglomerative . Divisive] **Community Detection**

Network-Centric | [Agglomerative . Divisive] Community Detection



Hierarchical Clustering

Hierarchical Clustering - procedure

1. Build a similarity matrix for the network

2. *Similarity matrix*: how similar two nodes are to each other → we need to determine from the adjacency matrix

3. Hierarchical clustering iteratively identifies groups of nodes with high similarity, following one of two distinct strategies:

   Agglomerative algorithms merge nodes and communities with high similarity.

   Divisive algorithms split communities by removing links that connect nodes with low similarity.

4. *Hierarchical tree* or *dendrogram*: visualize the history of the merging or splitting process the algorithm follows. Horizontal cuts of this tree offer various community partitions.

Network-Centric | [Agglomerative] Community Detection

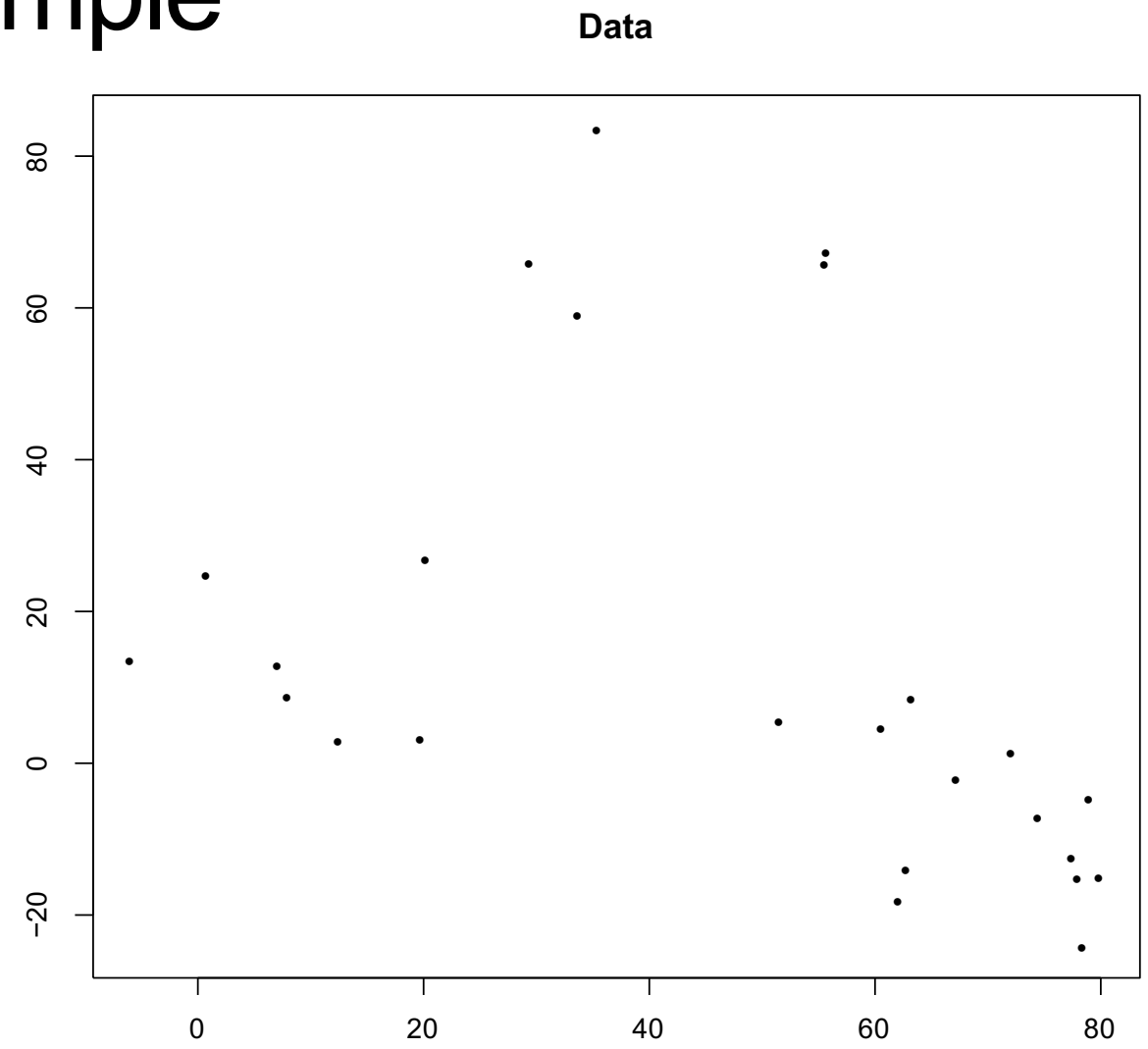
Similarity based vertex clustering:

- Define similarity measure between vertices based on network structure
  - Jaccard similarity
  - Cosine similarity
  - Pearson correlation
  - Eucledian distance (dissimilarity)
- Calculate similarity between all pairs of vertices in the graph (similarity matrix)
- Group together vertices with high similarities

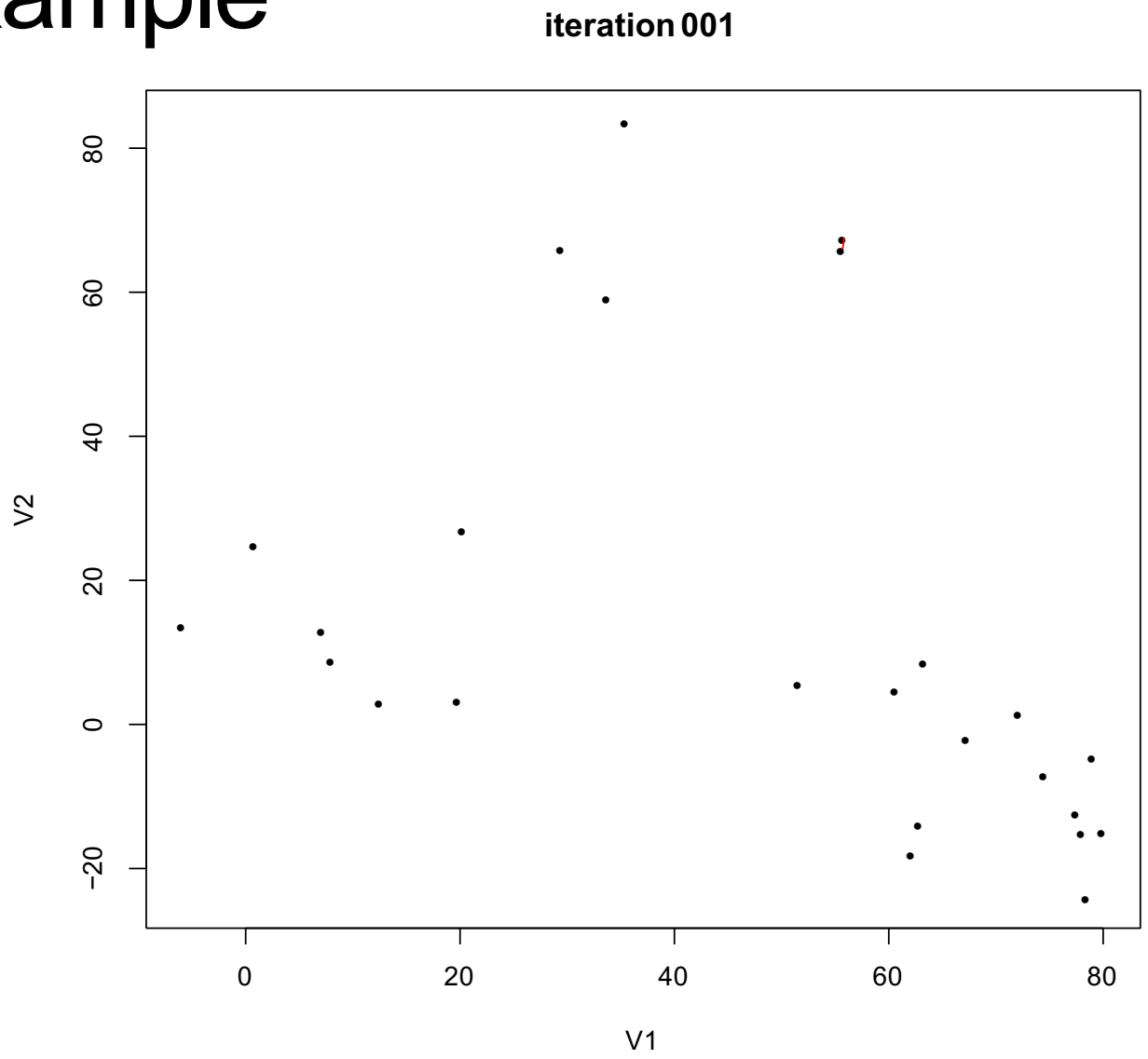
Pseudocode

1. Assign each node to its own cluster
2. Find the cluster pair with highest similarity and join them together into a cluster
3. Compute new similarities between new joined cluster and others
4. Go to step 2 until all nodes form a single cluster

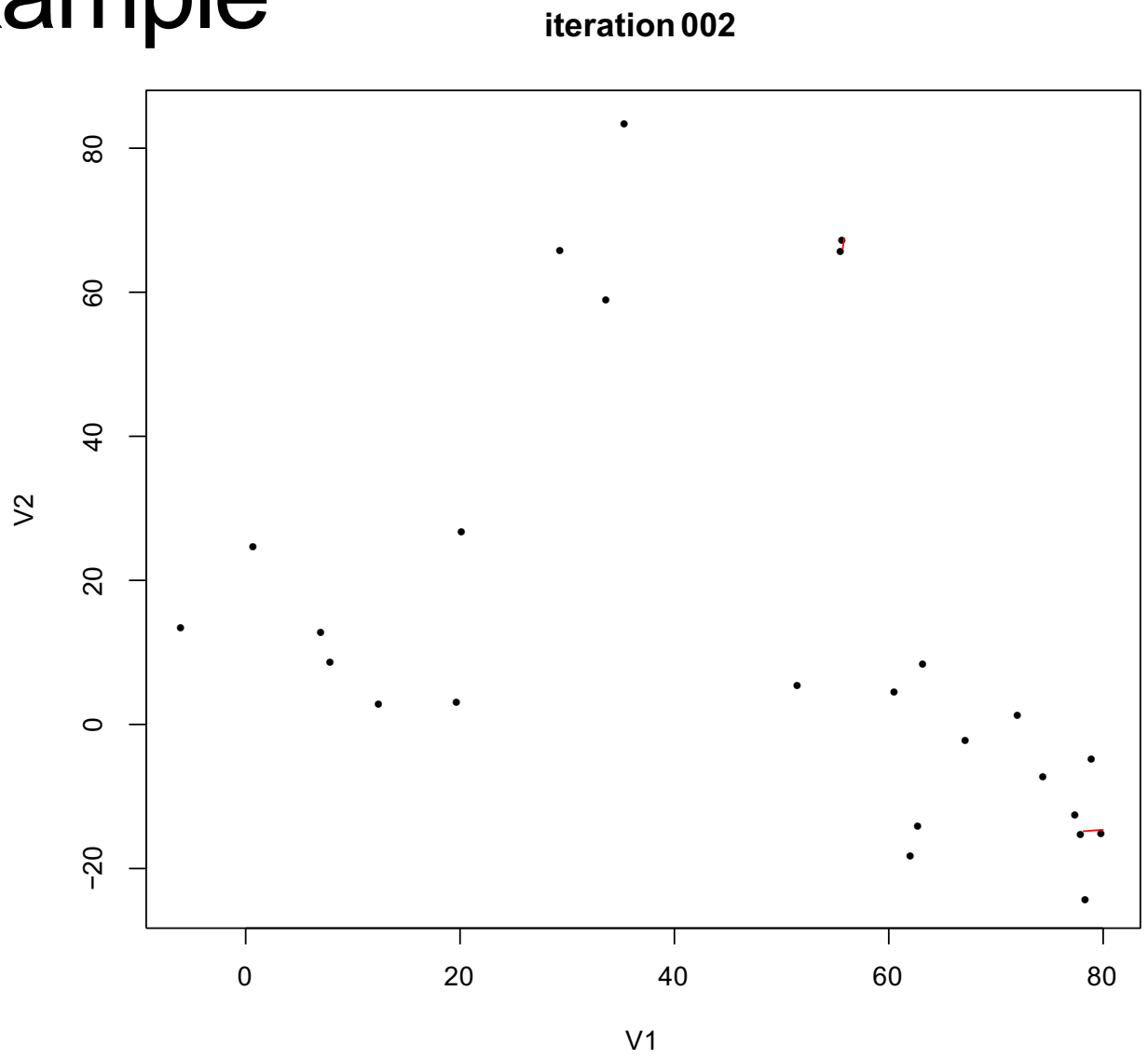
Example



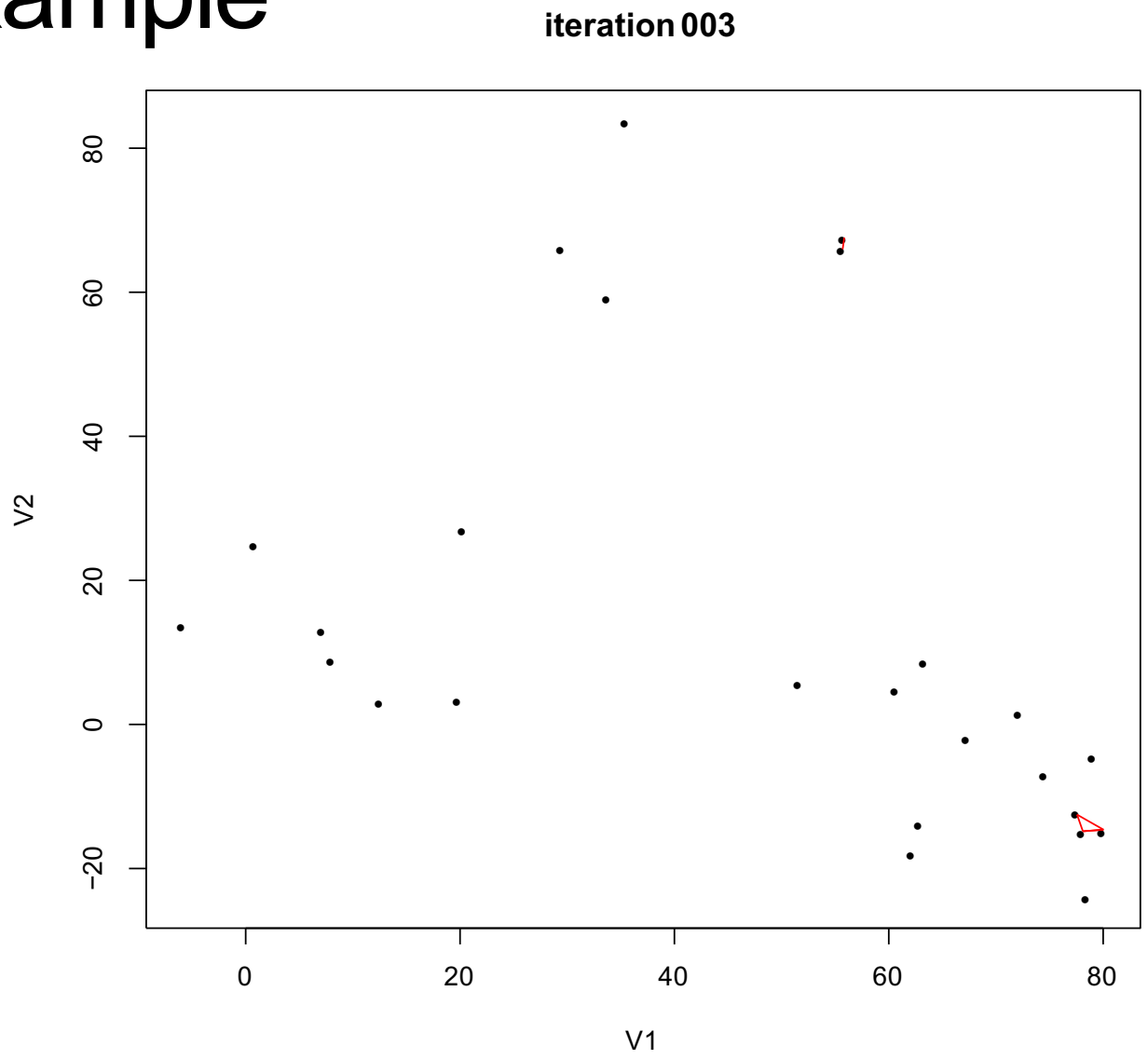
Example



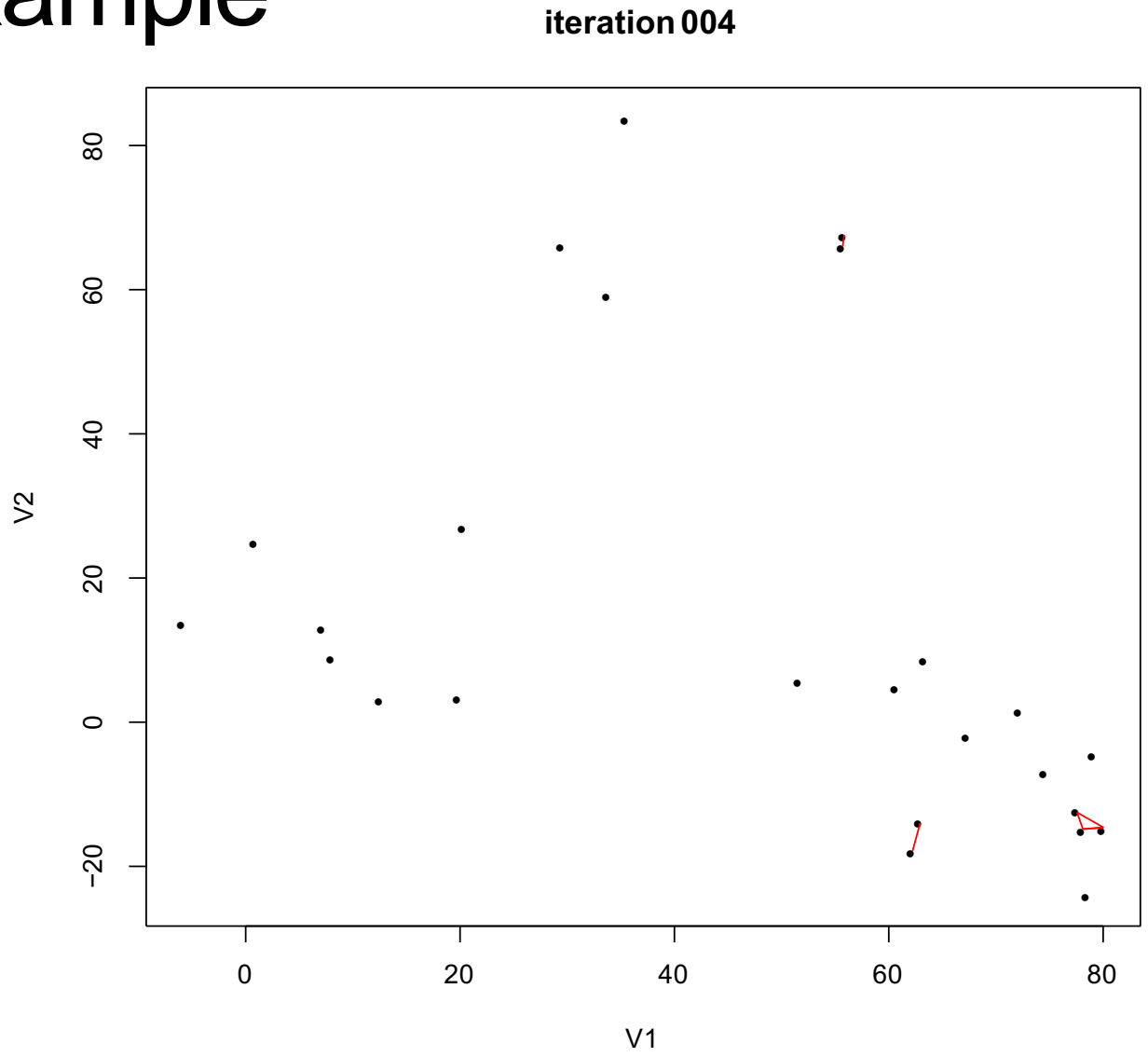
Example



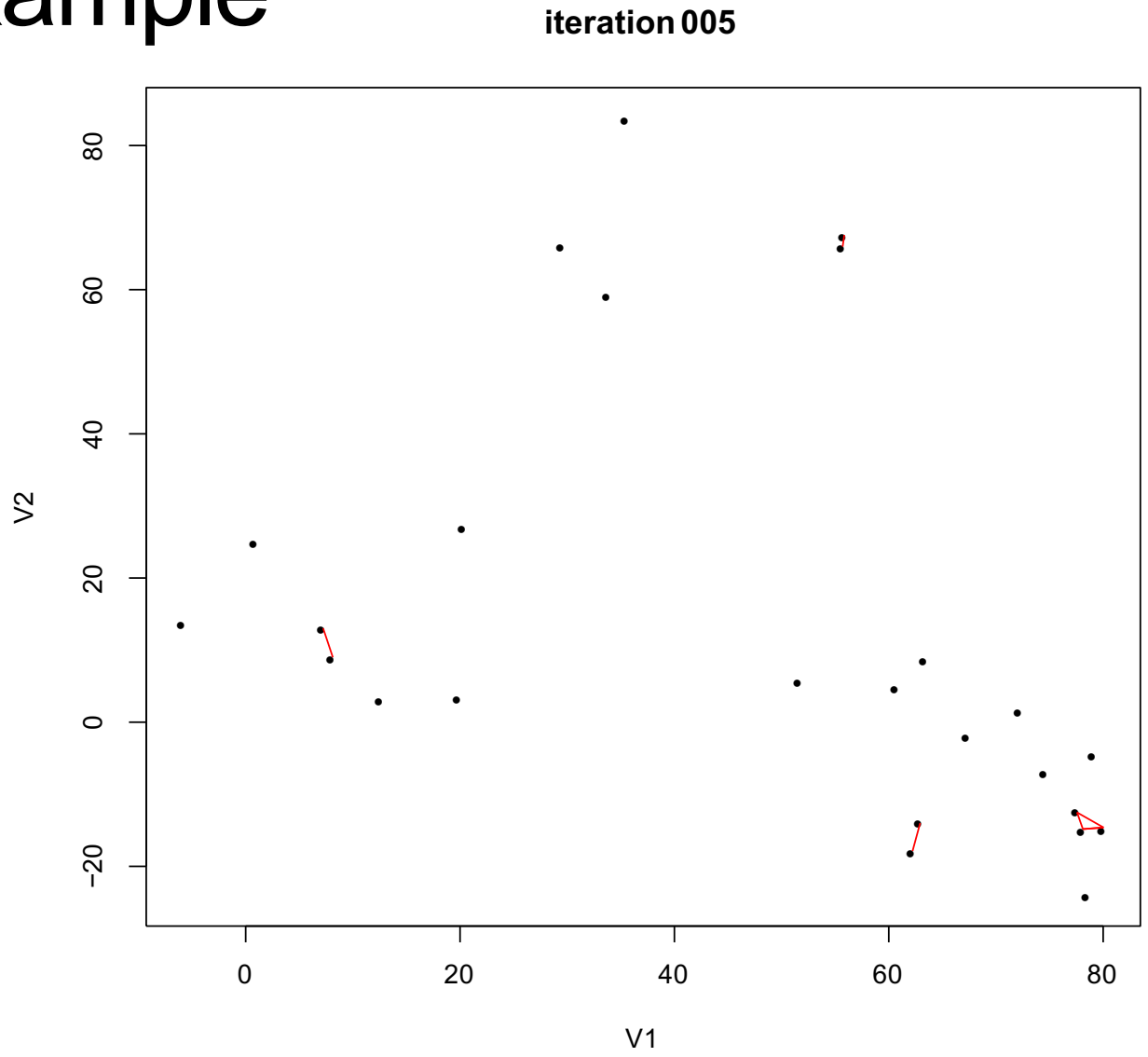
Example



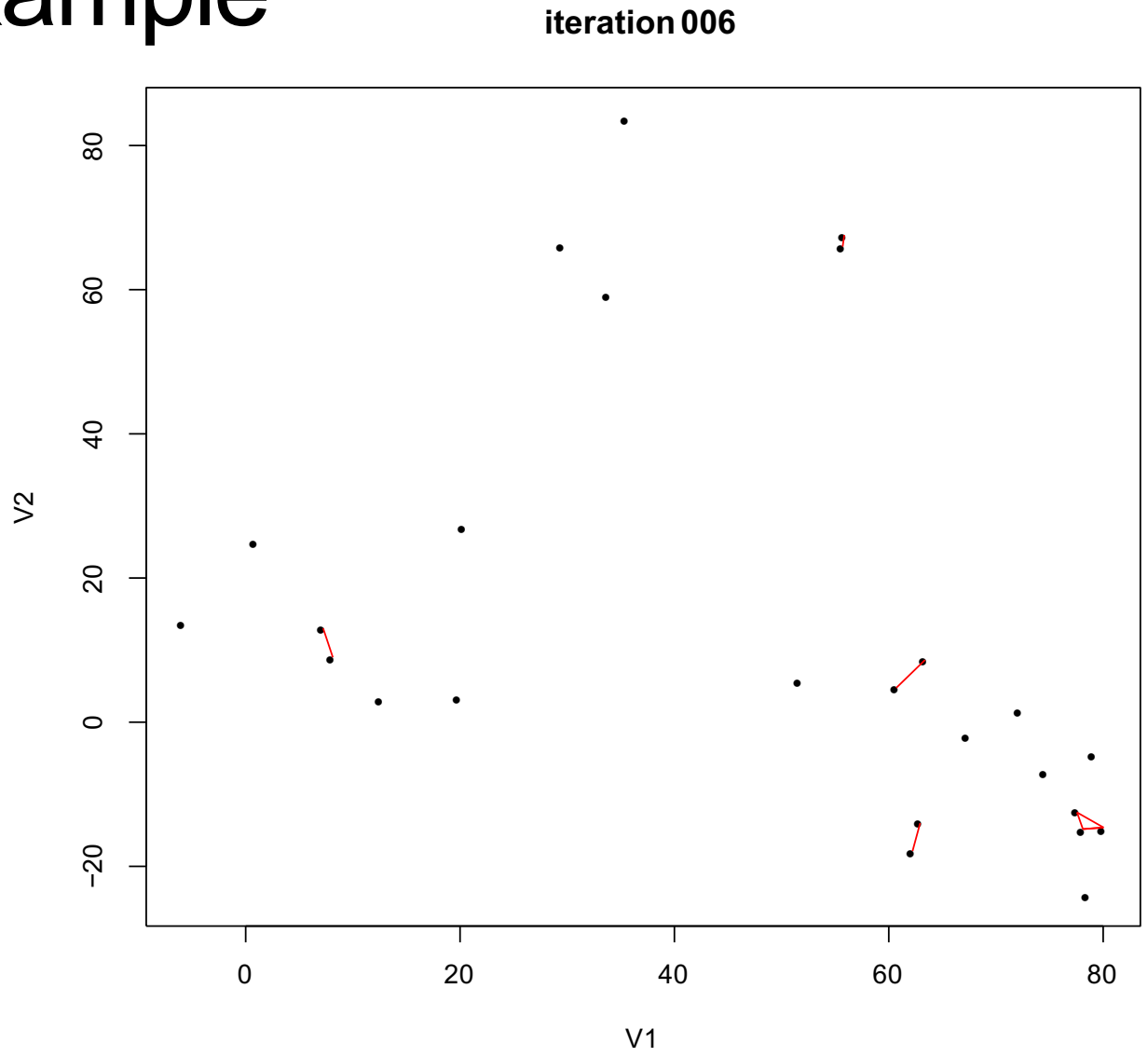
Example



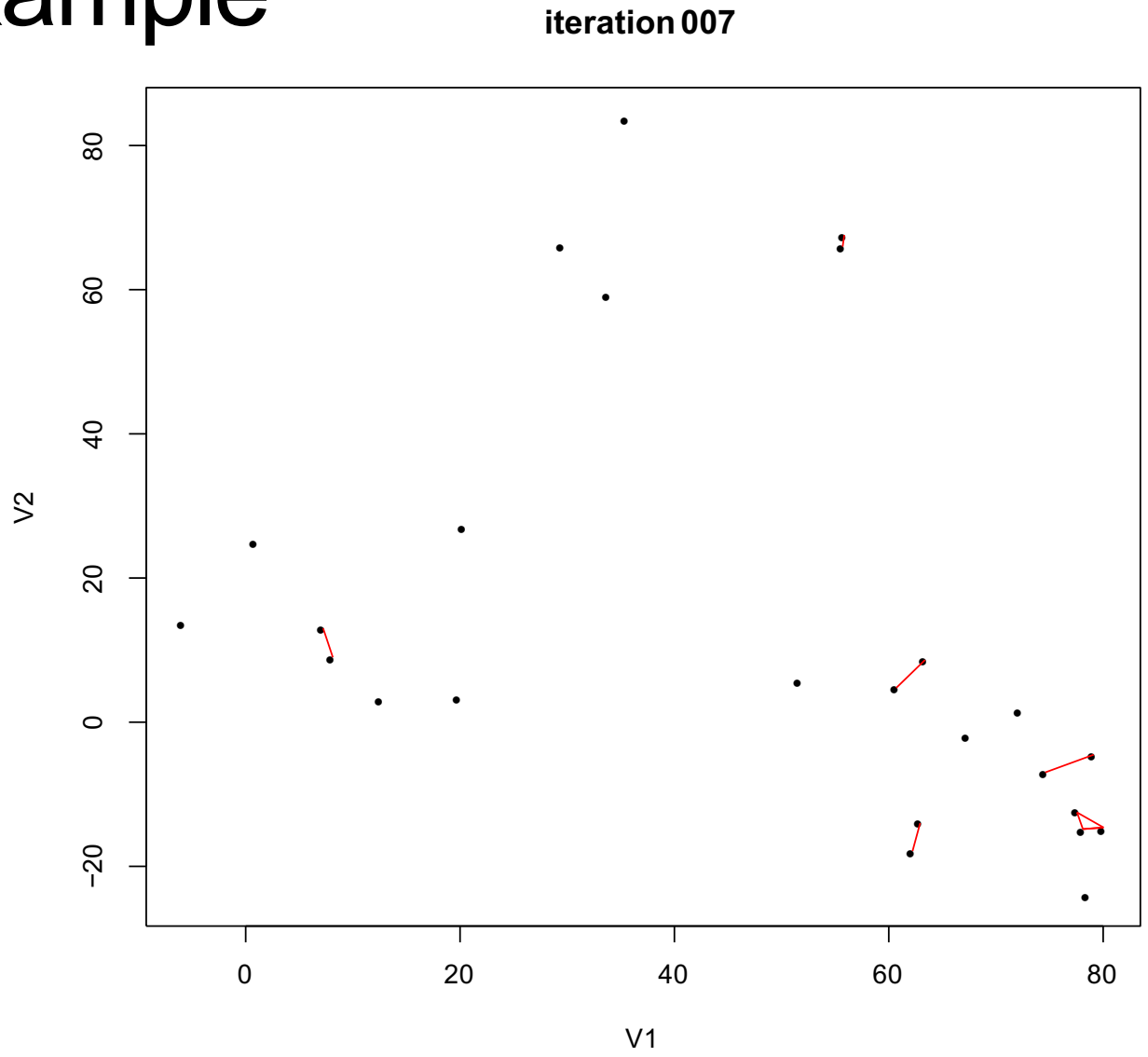
Example



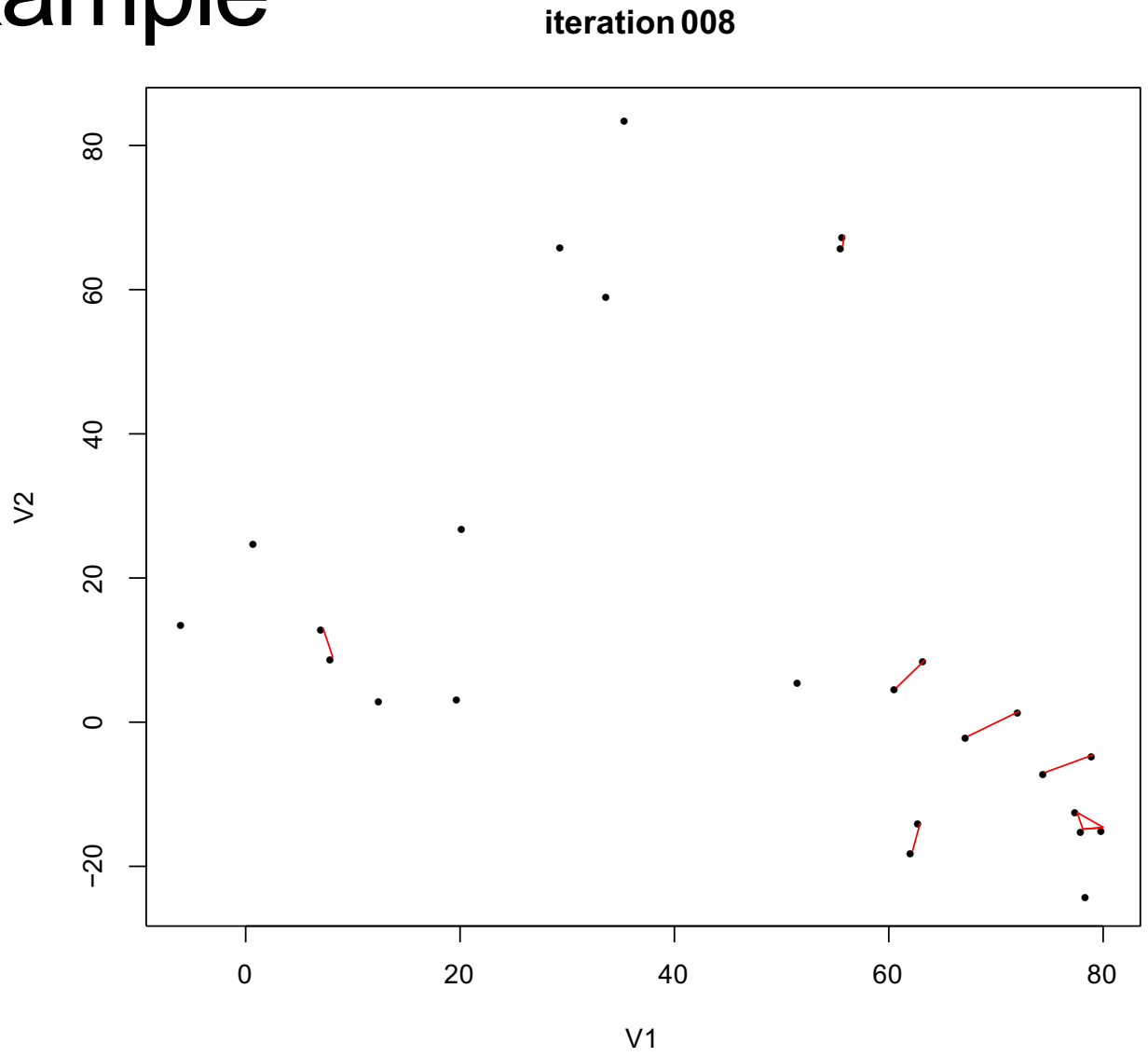
Example



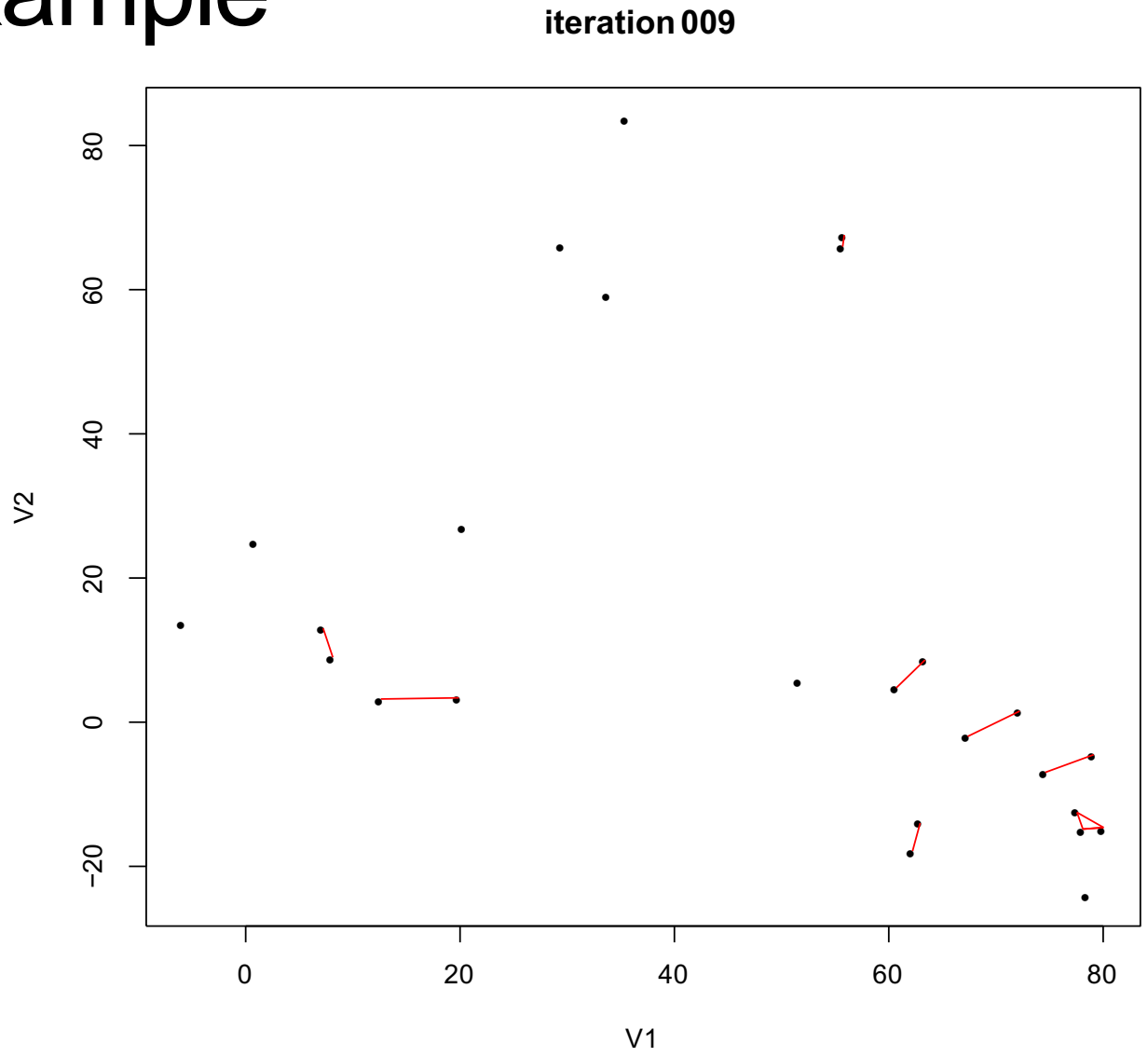
Example



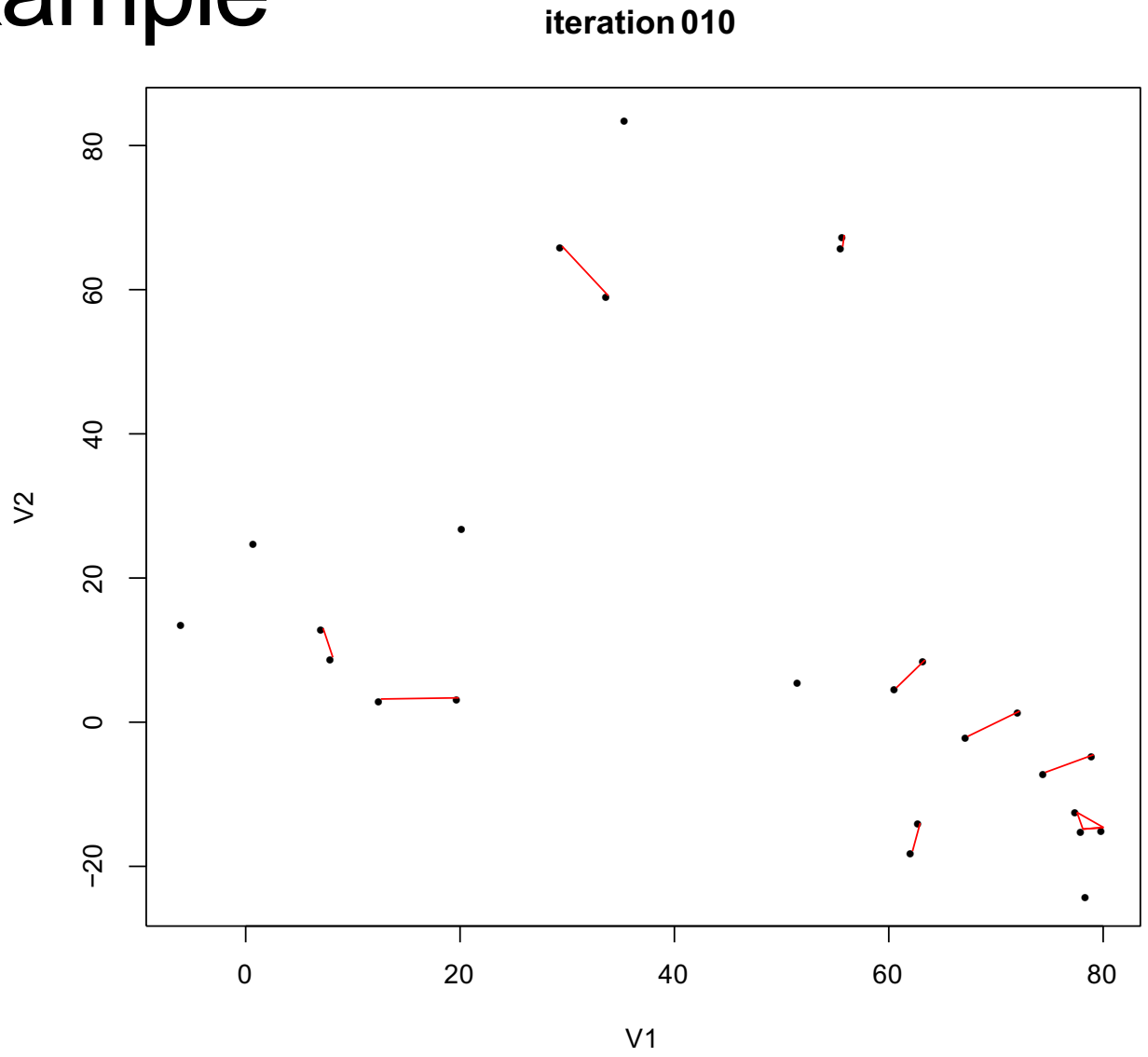
Example



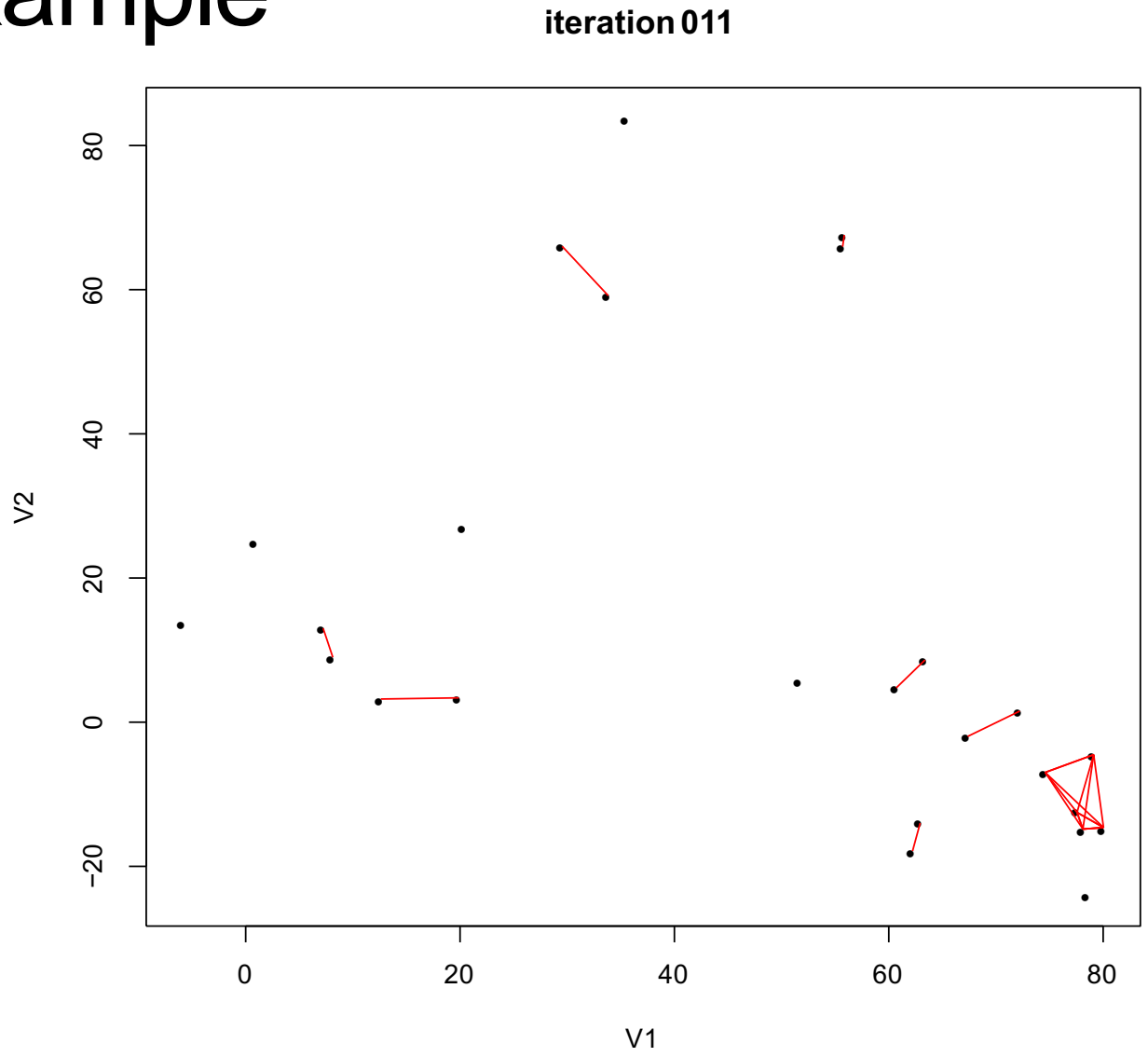
Example



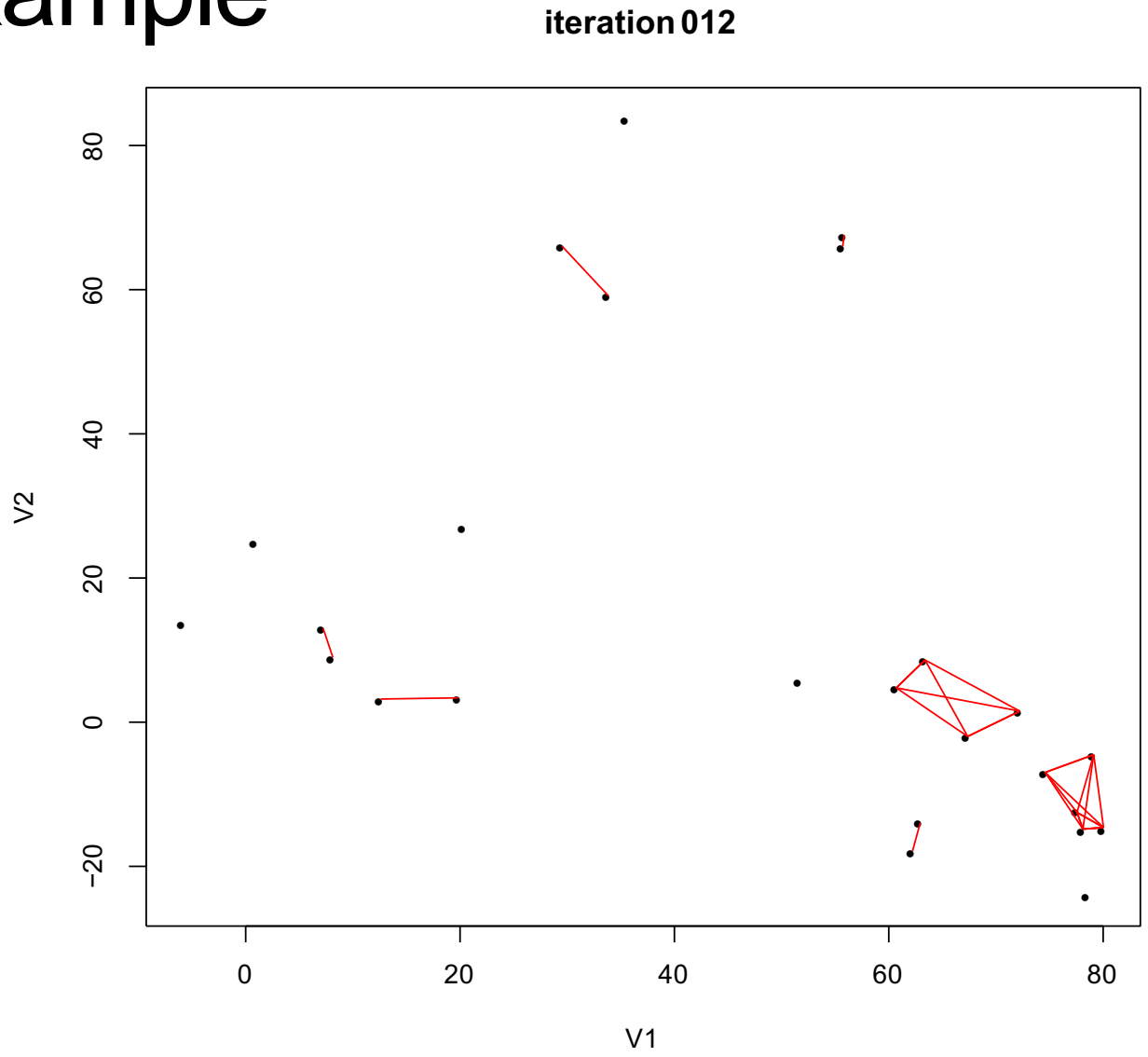
Example



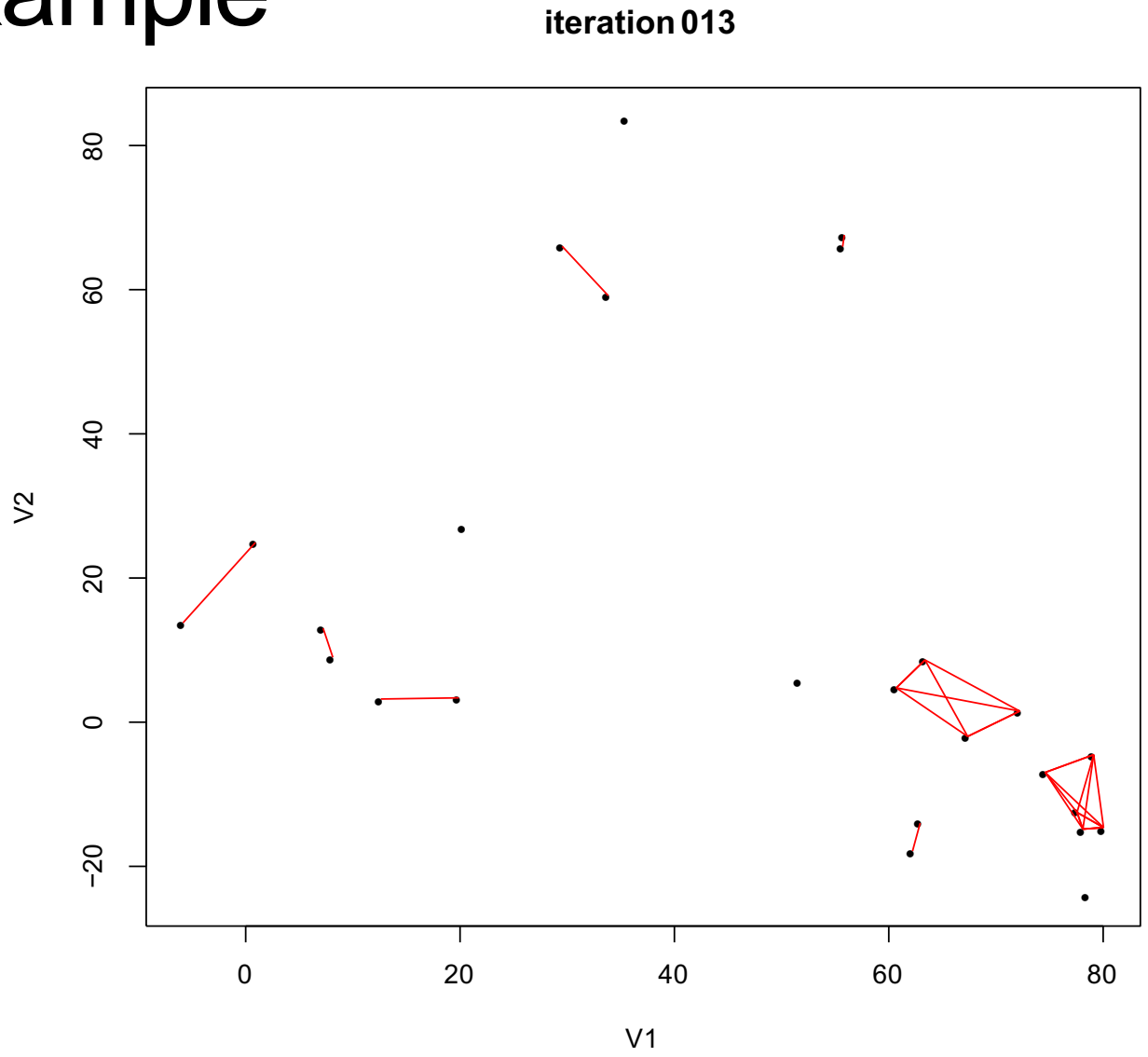
Example



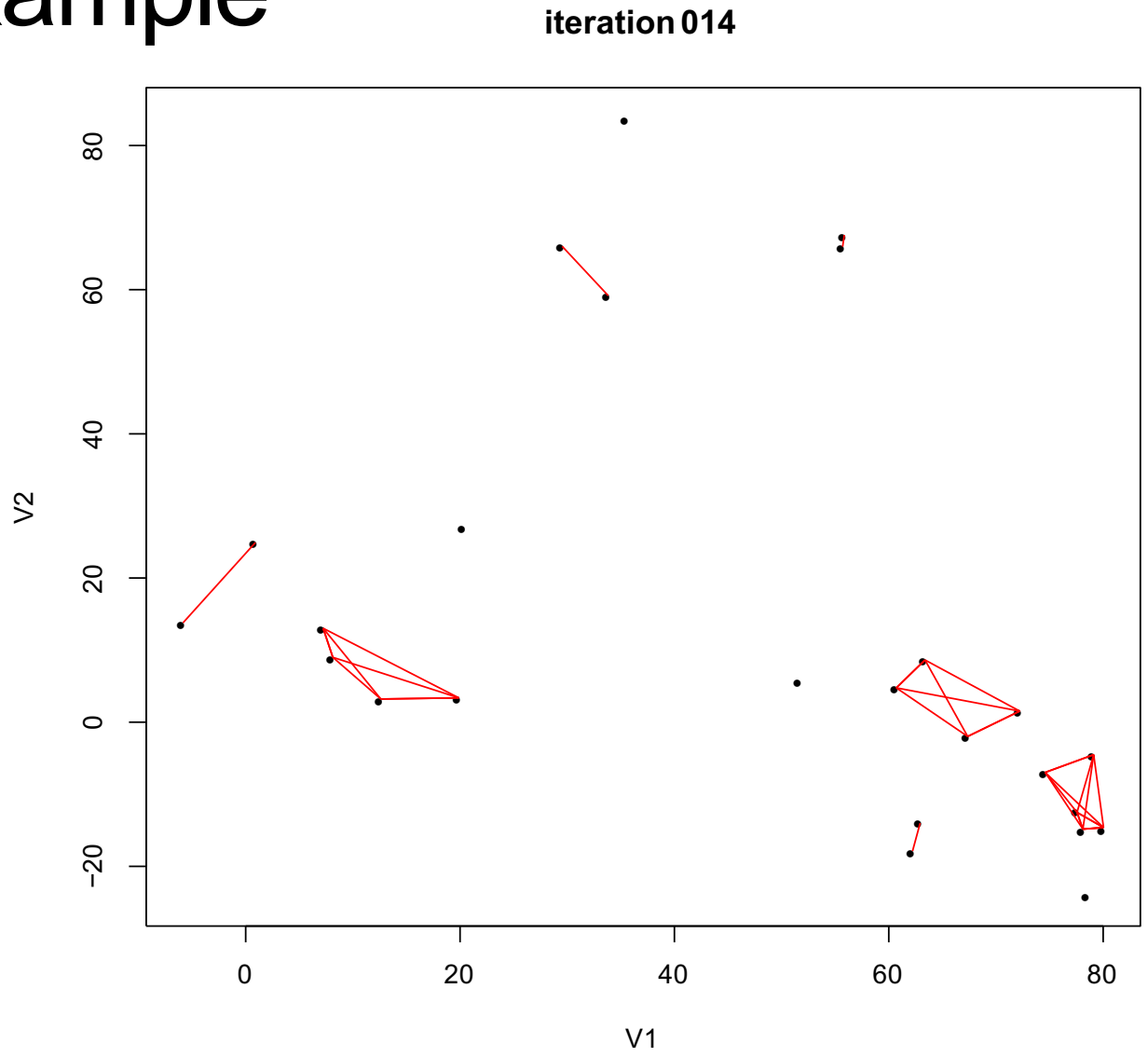
Example



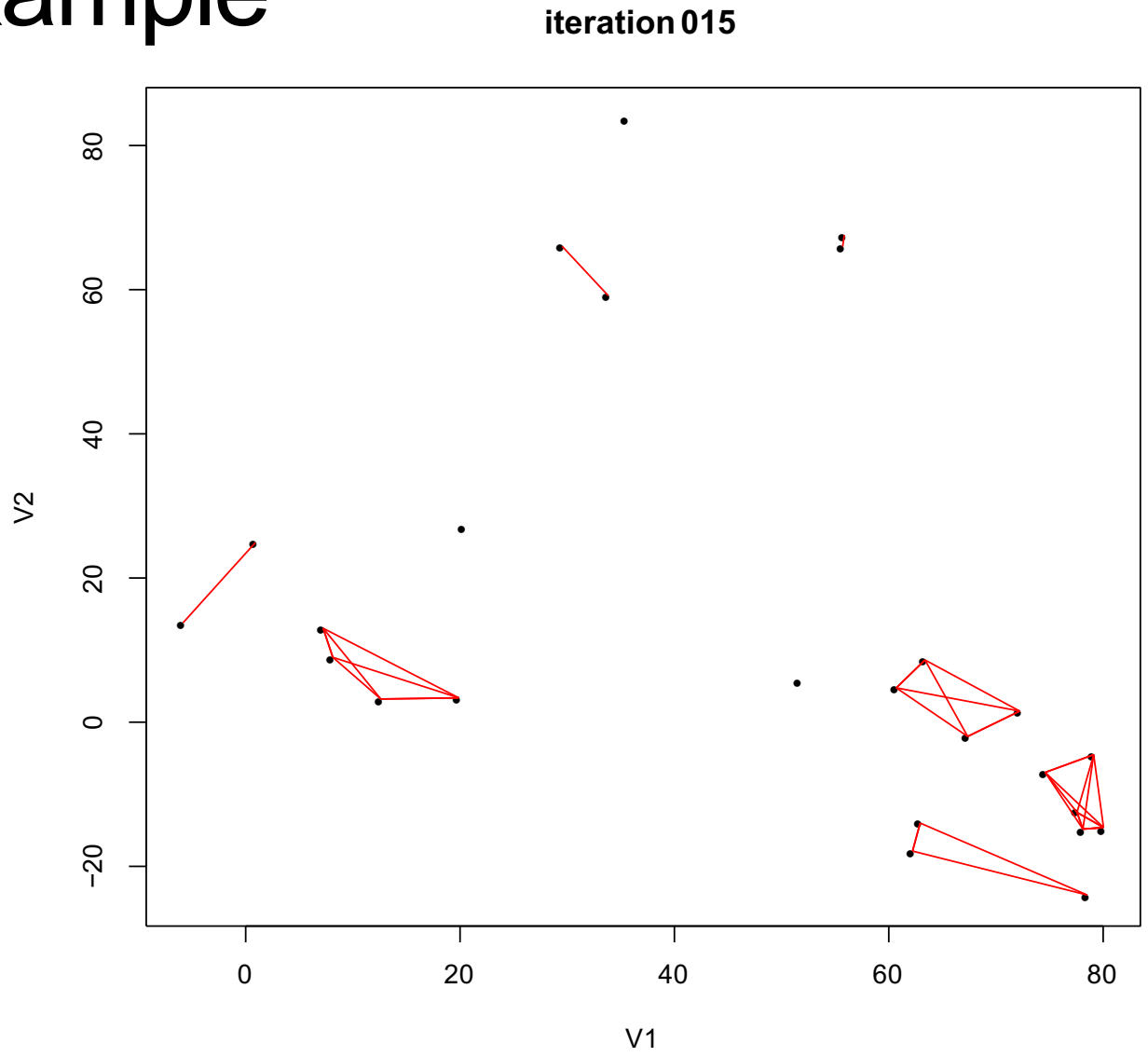
Example



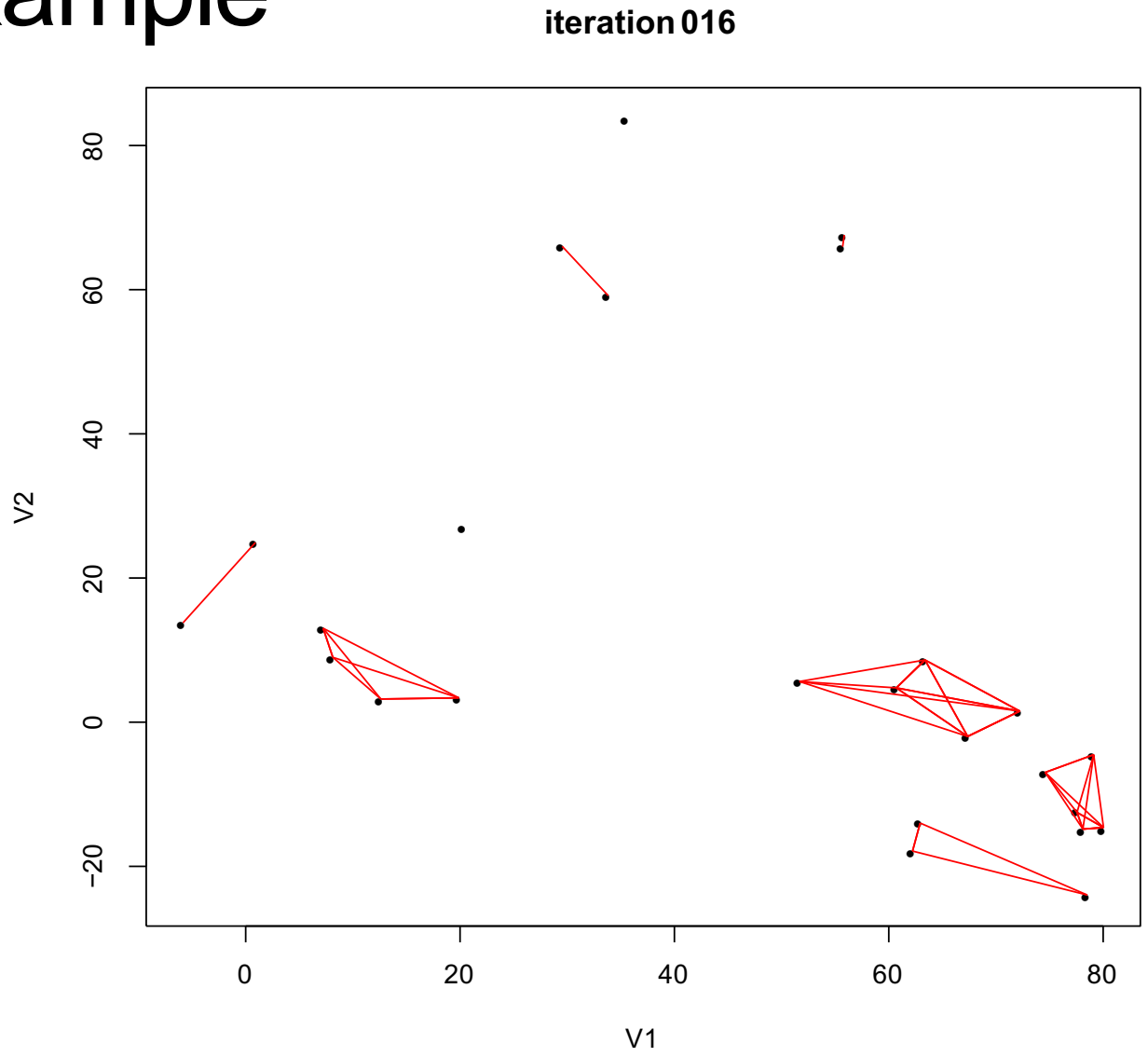
Example



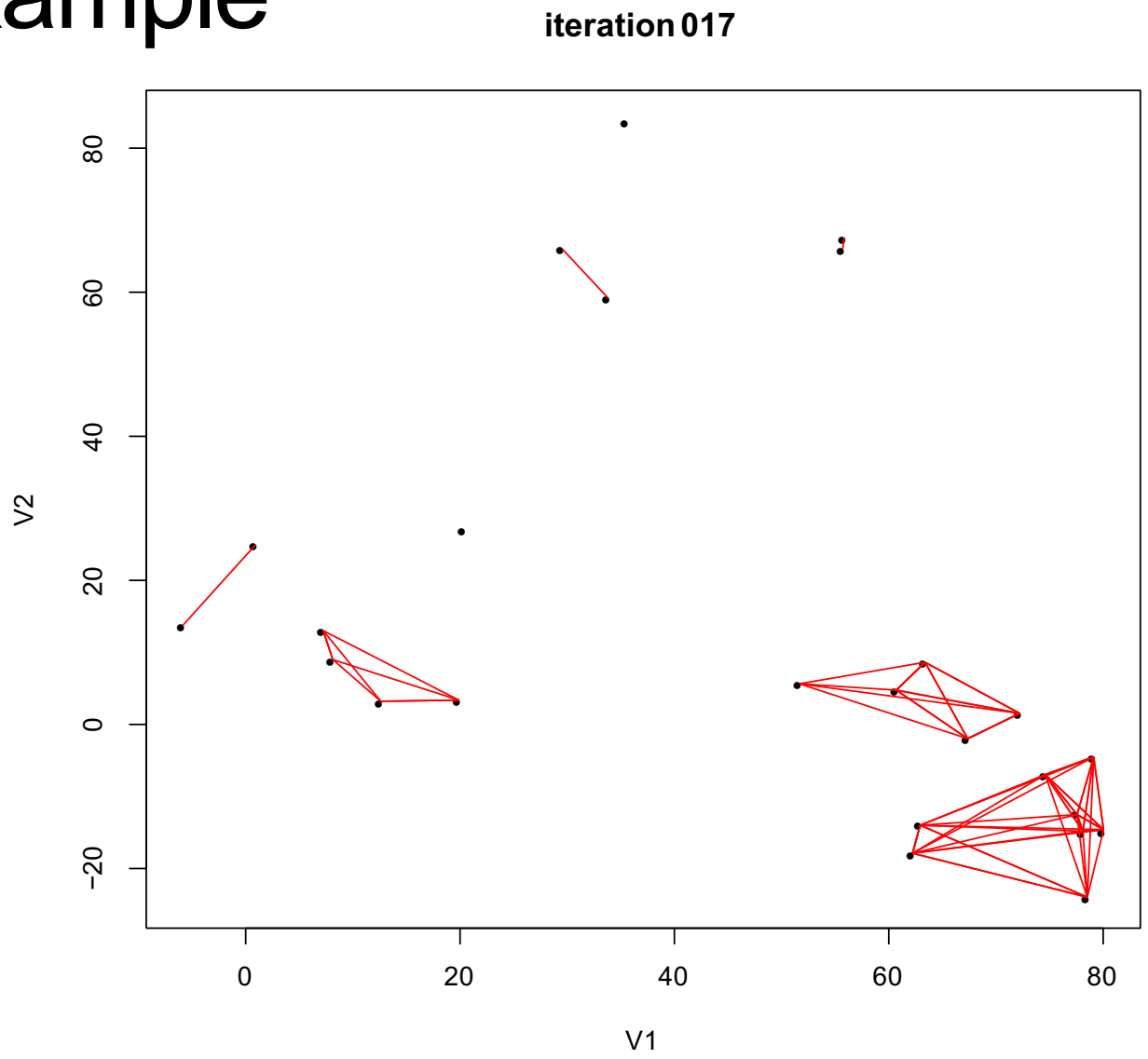
Example



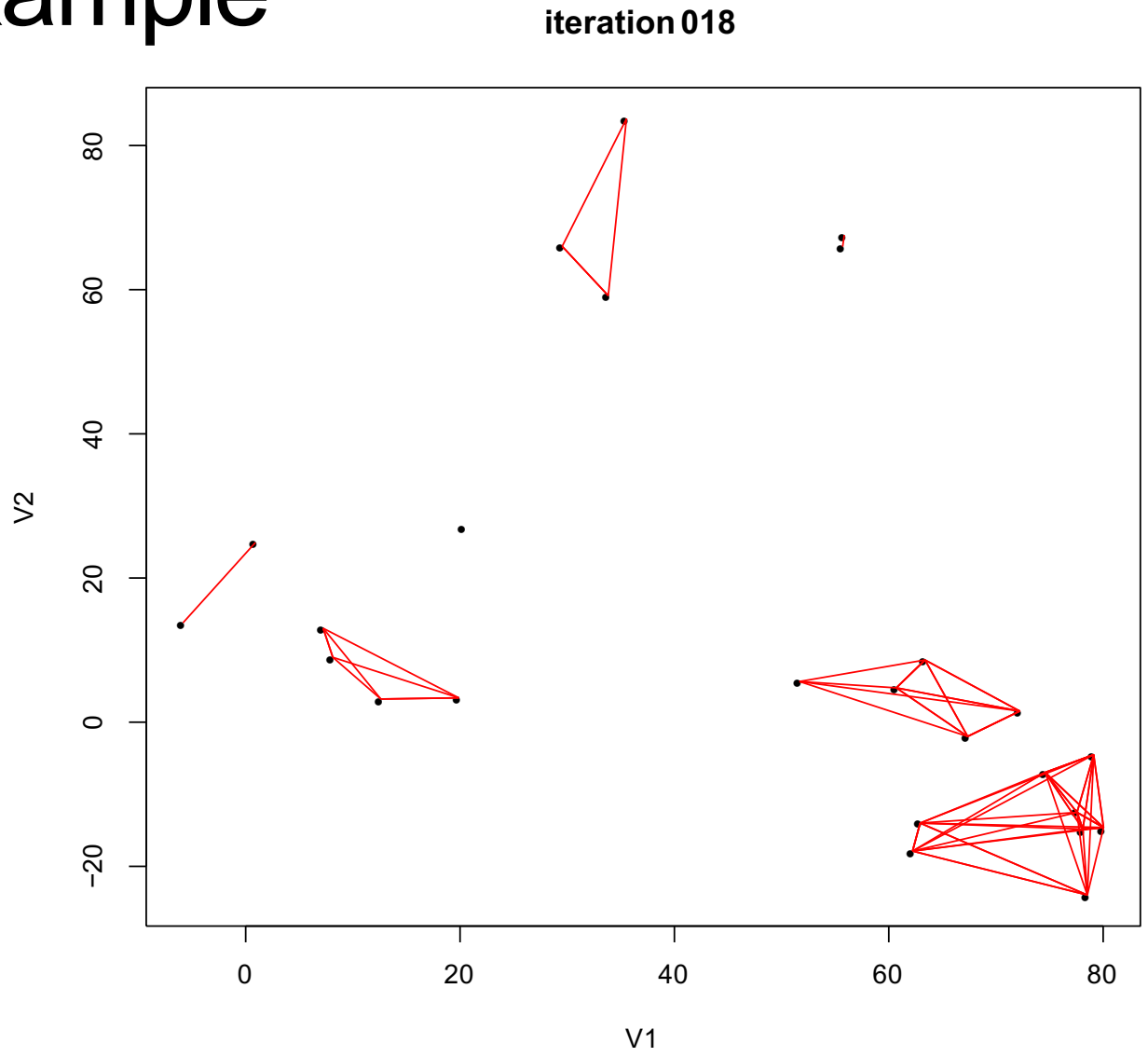
Example



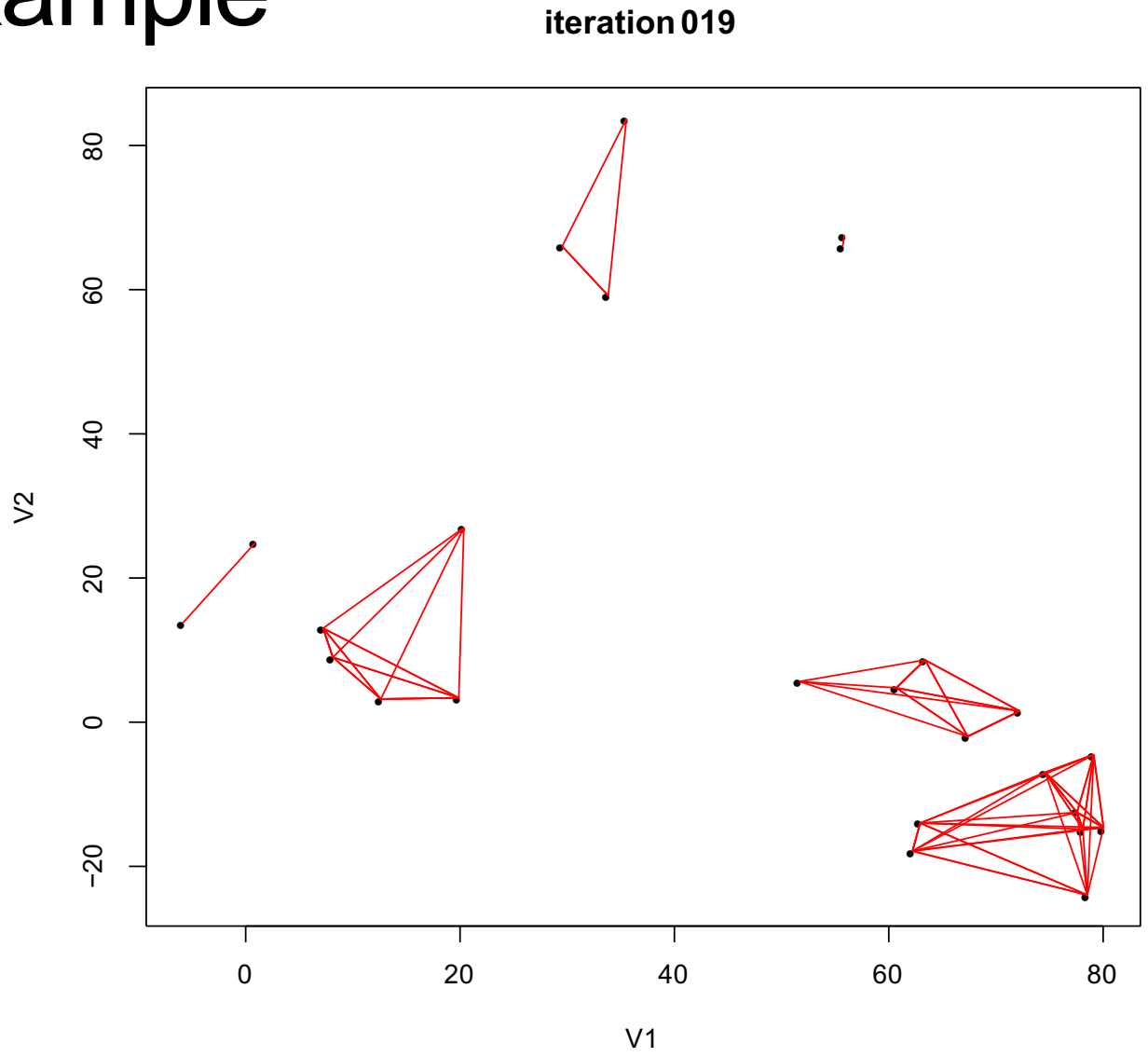
Example



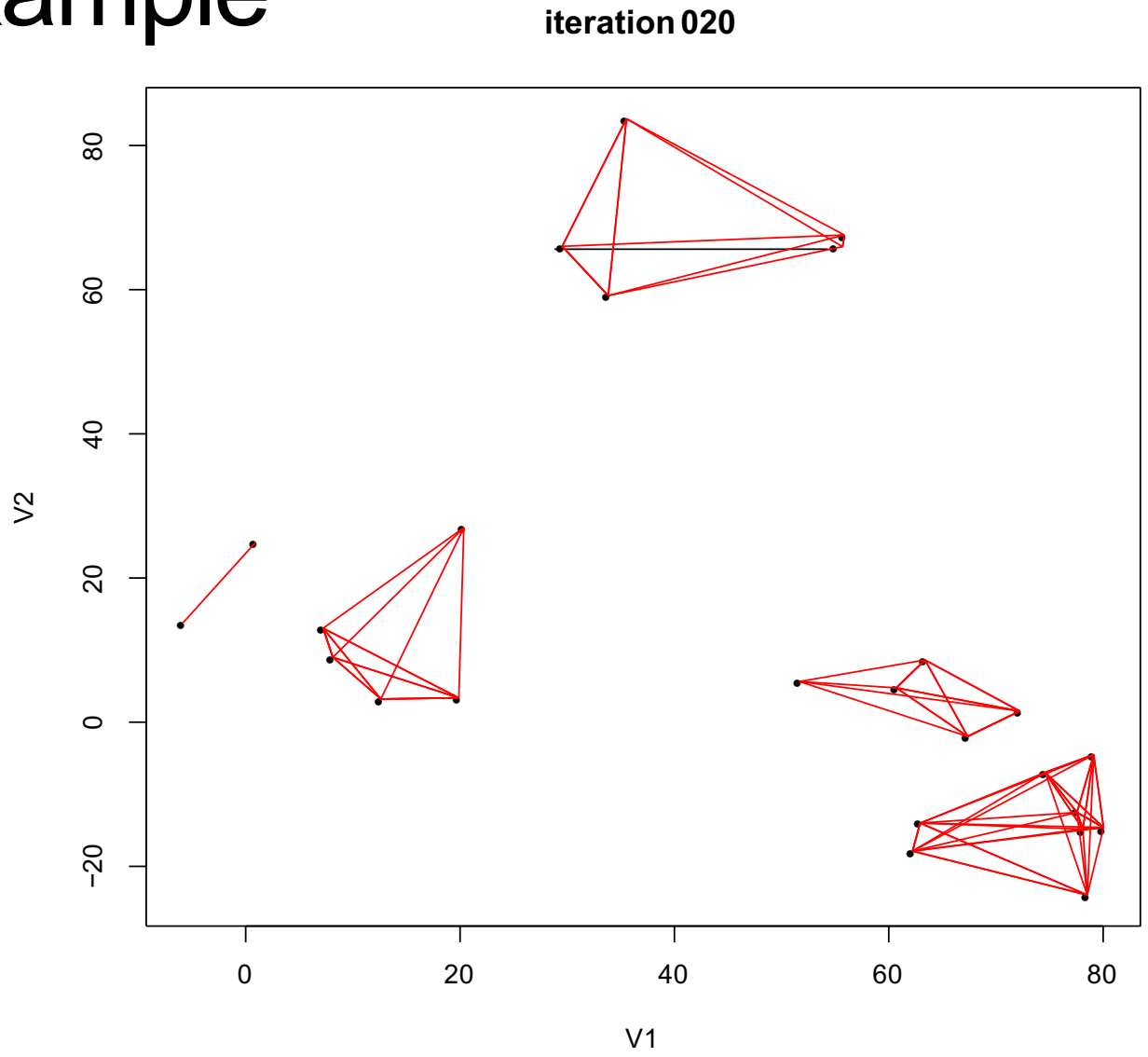
Example



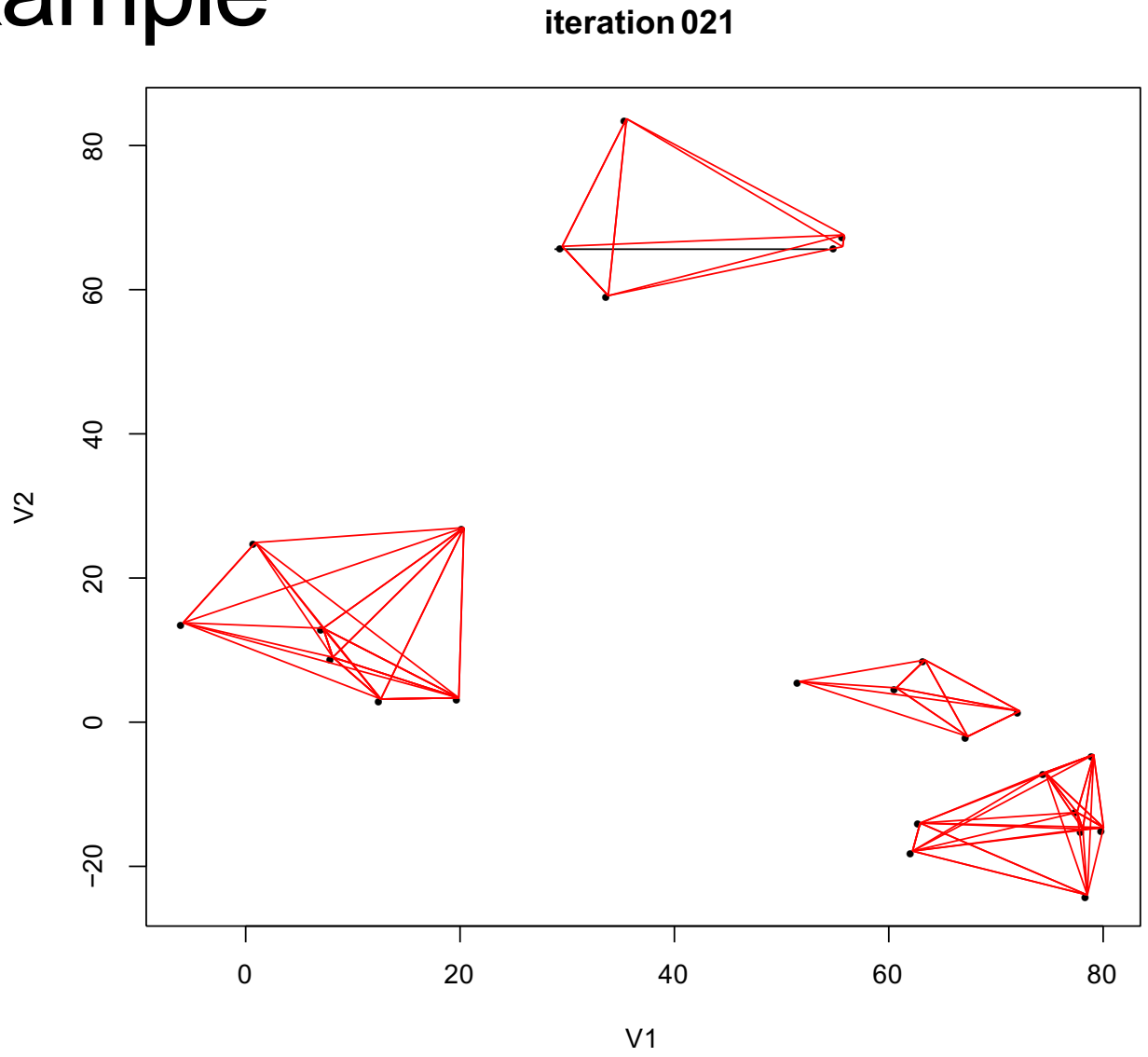
Example



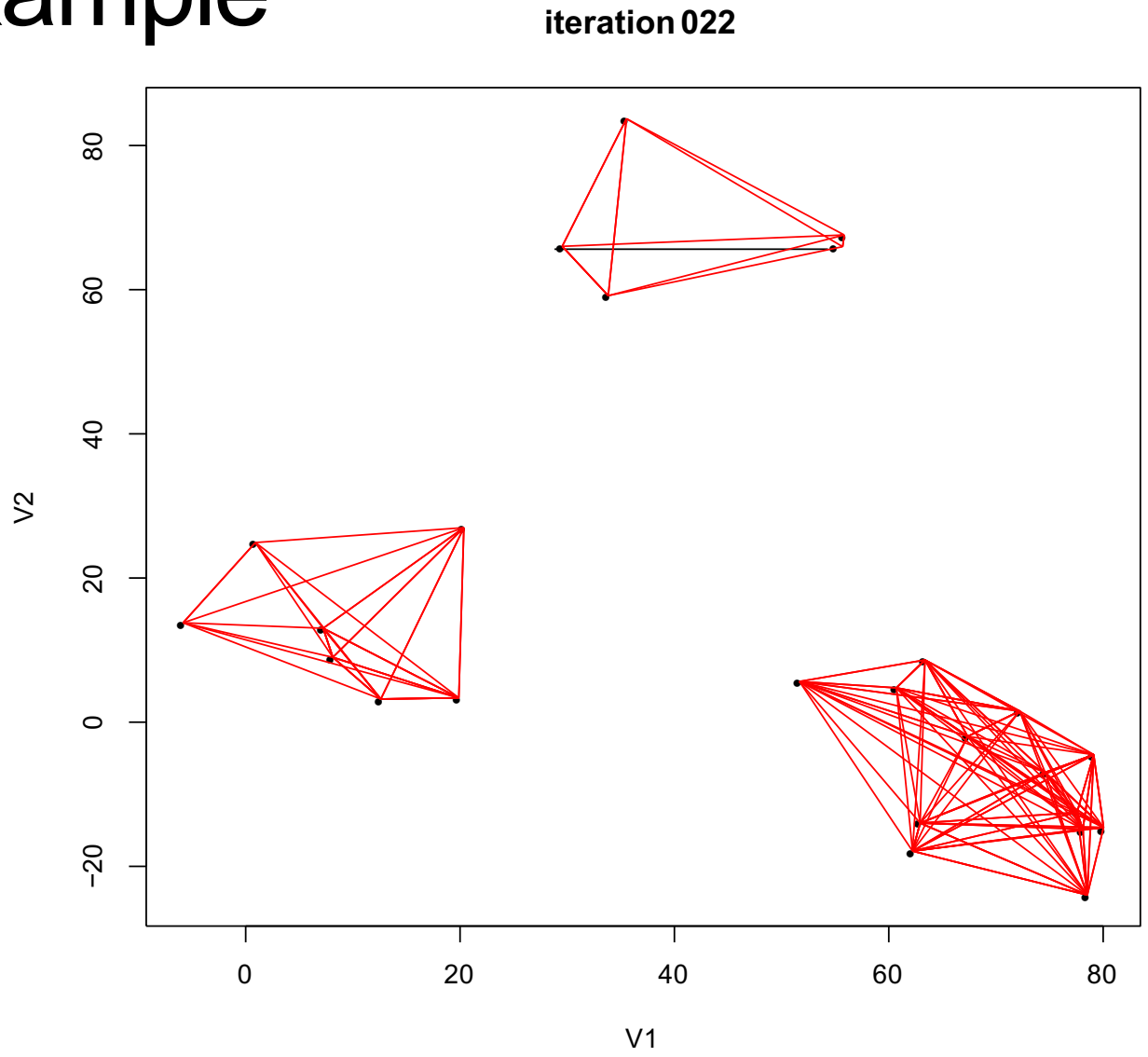
Example



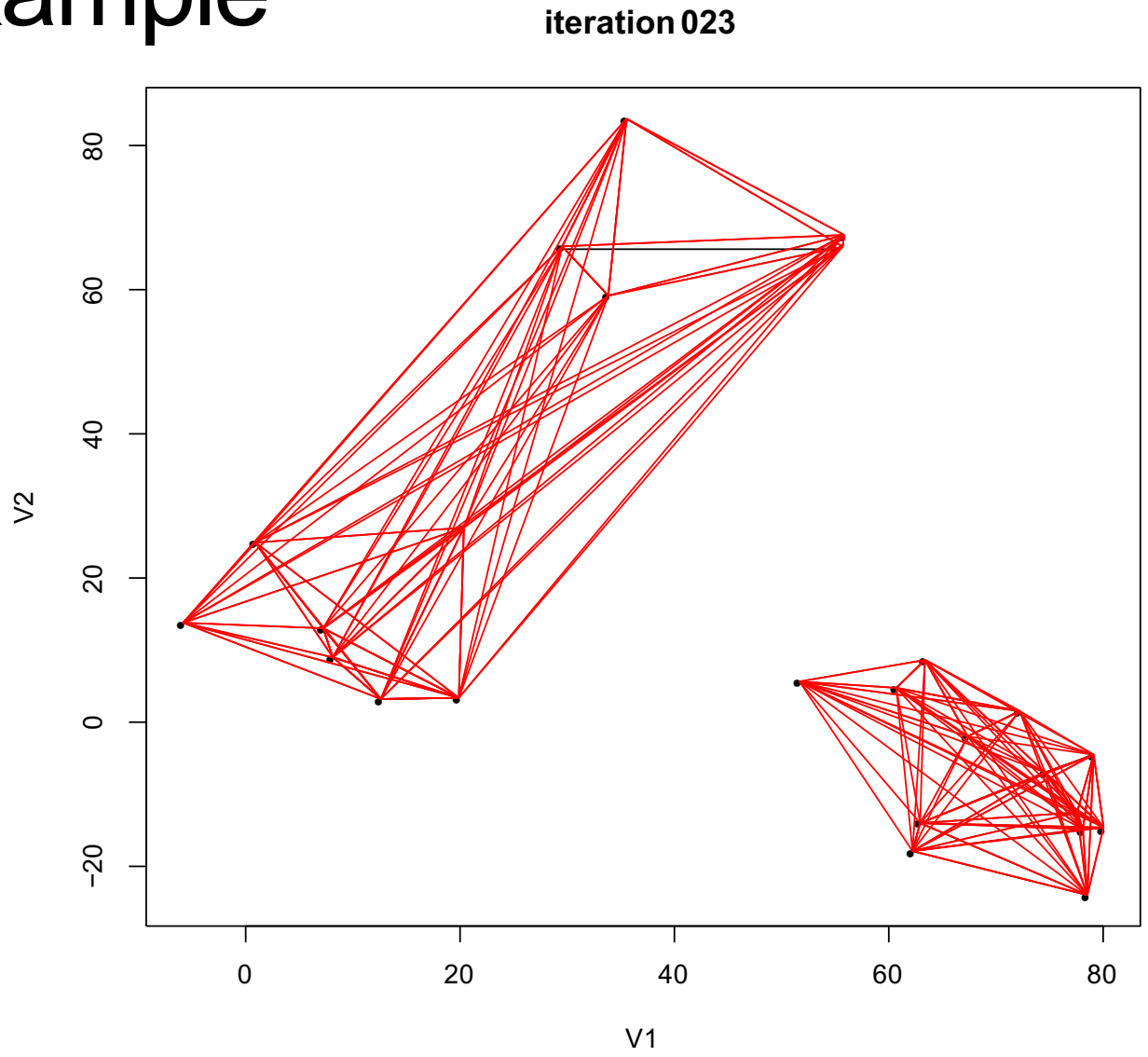
Example



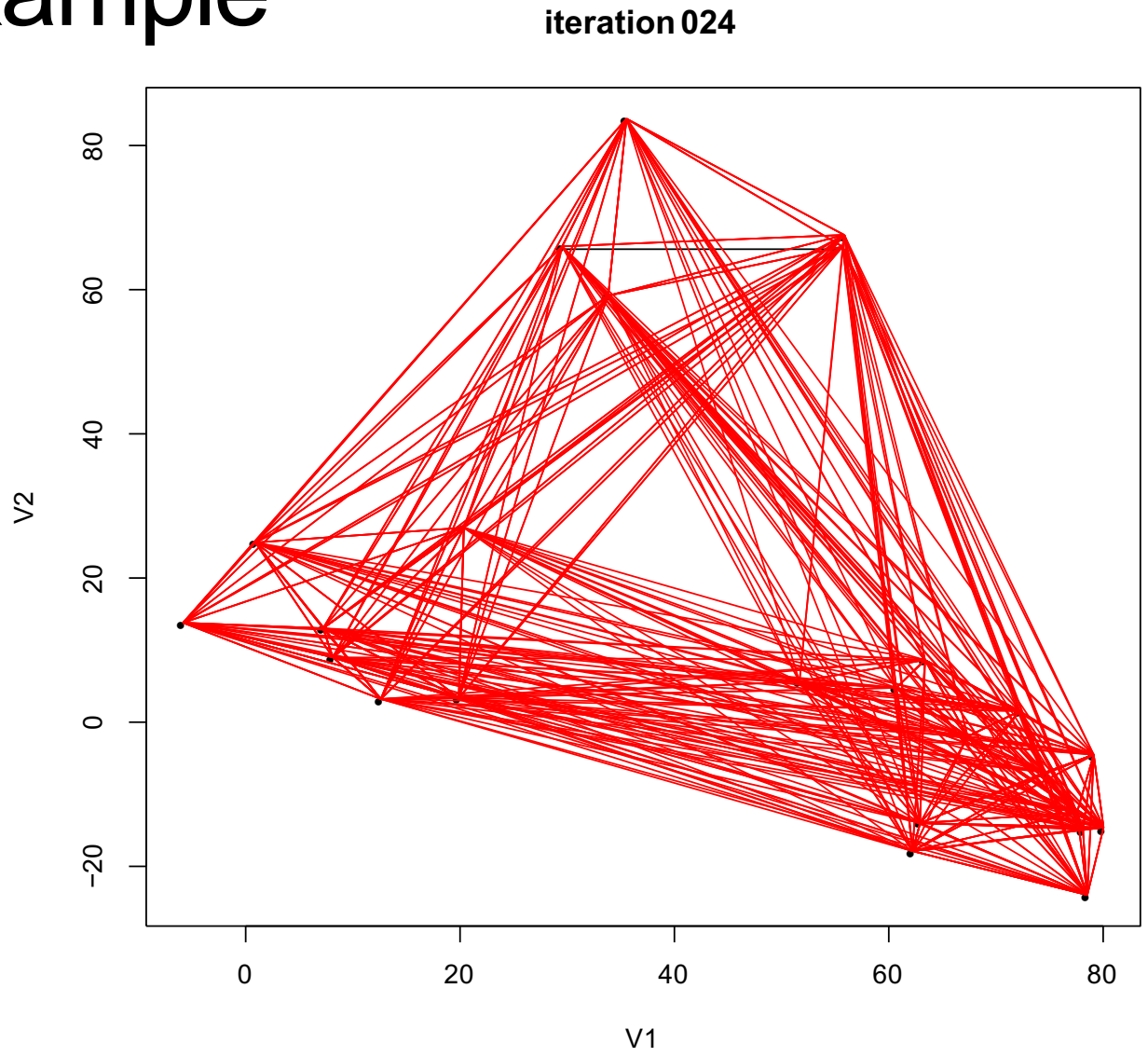
Example



Example

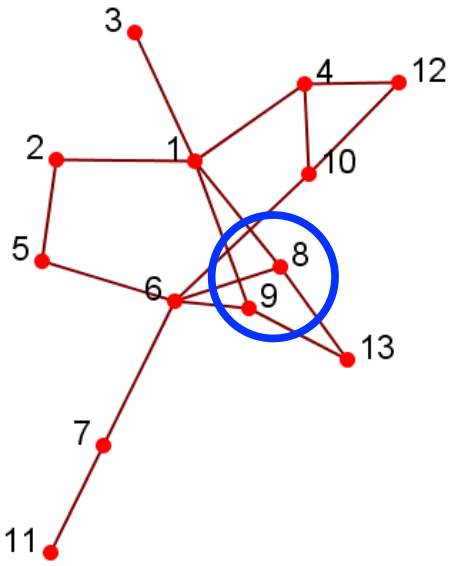


Example



Similarity Measures | structural equivalence or vector similarity

- Node similarity is defined by how similar their interaction patterns are
- Two nodes are structurally equivalent if they connect to the same set of actors
  - e.g., nodes 8 and 9 are structurally equivalent
- Groups are defined over equivalent nodes
  - Too strict
  - Rarely occur in a large-scale
  - Relaxed equivalence class is difficult to compute
- In practice, use vector similarity
  - e.g., cosine similarity, Jaccard similarity

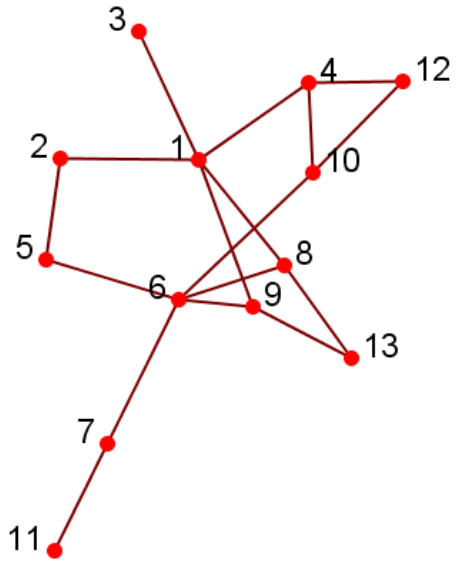


Similarity Measures | structural equivalence or vector similarity (Cosine v Jaccard)

	1	2	3	4	5	6	7	8	9	10	11	12	13
a vector → 5		1				1							
structurally equivalent 8	1					1							1
structurally equivalent 9	1					1							1

Cosine Similarity: $$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}.$$

$$sim(5,8) = \frac{1}{\sqrt{2} \times \sqrt{3}} = \frac{1}{\sqrt{6}}$$



Jaccard Similarity: $$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

$$J(5,8) = \frac{|\{6\}|}{|\{1,2,6,13\}|} = 1/4$$

Similarity Measures for nodes | euclidean distance & pearson correlation

Euclidean distance: (or rather Hamming distance since A is binary)

$$d_{ij} = \sum_k (A_{ik} - A_{jk})^2$$

Normalized Euclidean distance:[2]

$$d_{ij} = \frac{\sum_k (A_{ik} - A_{jk})^2}{k_i + k_j} = 1 - 2\frac{n_{ij}}{k_i + k_j}$$

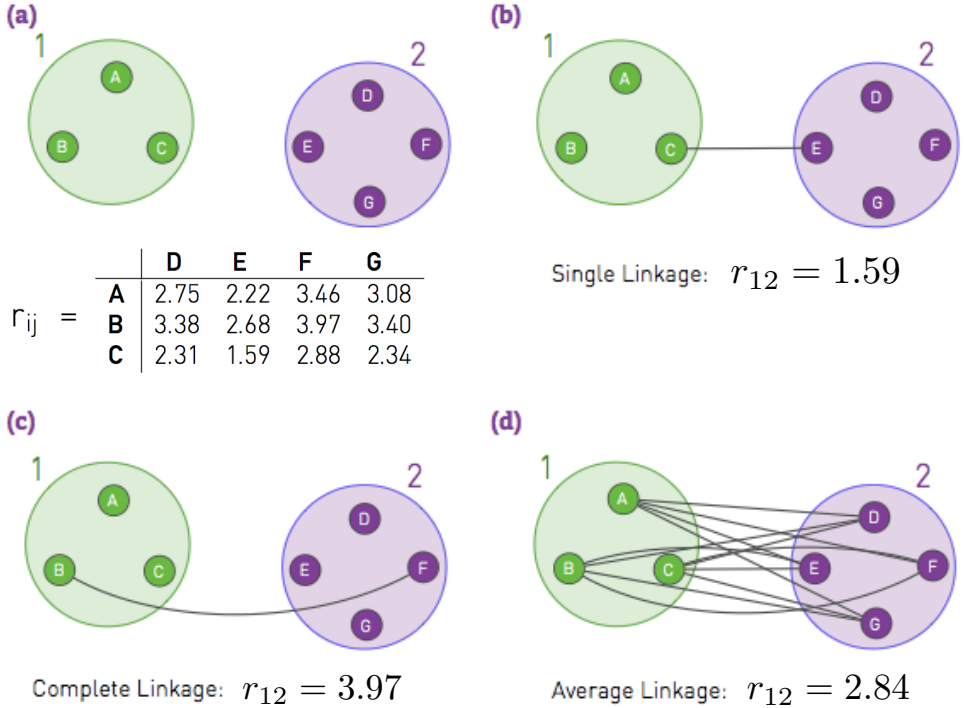
Pearson correlation coefficient

$$r_{ij} = \frac{cov(A_i, A_j)}{\sigma_i \sigma_j} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i \sigma_j}$$

where $\mu_i = \frac{1}{n}\sum_k A_{ik}$ and $\sigma_i = \sqrt{\frac{1}{n}\sum_k (A_{ik} - \mu_i)^2}$

Decide GROUP SIMILARITY| Agglomerative Hierarchical clustering

- Single linkage: $s_{XY} = \min_{x \in X, y \in Y} s_{xy}$
- Complete linkage: $s_{XY} = \max_{x \in X, y \in Y} s_{xy}$
- Average linkage: $s_{XY} = \frac{\sum_{x \in X, y \in Y} s_{xy}}{|X| \times |Y|}$



$\frac{1}{x_{ij}} = r_{ij} =$

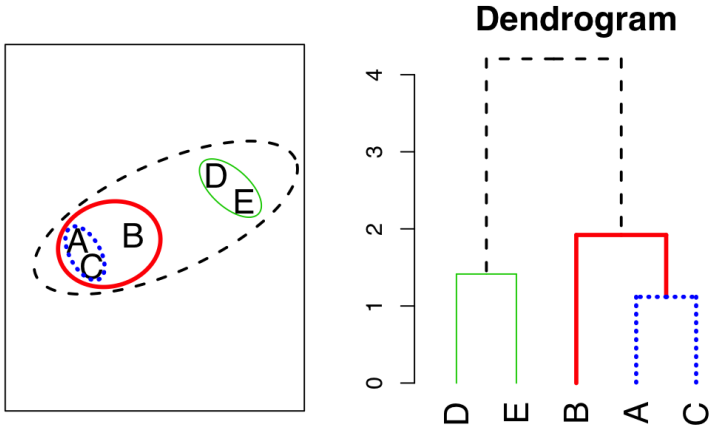
	D	E	F	G
A	2.75	2.22	3.46	3.08
B	3.38	2.68	3.97	3.40
C	2.31	1.59	2.88	2.34

Single linkage: similarity of two clusters is the similarity of their *most similar* or closest members; we only pay attention to the area where the two clusters come closest to each other – we're connecting a point to a nearby point. tends to produce long chains.
[only wants **one** point in the cluster to be close to another point in a different cluster]

Complete linkage: similarity of two clusters is the similarity of their *most dissimilar* members. chooses farthest elements in clusters.
[makes sure all points in two clusters are close to each other]

Clustering on Node Similarities | Agglomerative Hierarchical clustering

- Assign each vertex to a group of its own
- Find two groups with the highest similarity and join them in a single group
- Calculate similarity between groups:
  - single-linkage clustering (most similar in the group)
  - complete-linkage clustering (least similar in the group)
  - average-linkage clustering (mean similarity between groups)
- Repeat until all joined into single group



Johnson's Hierarchical Clustering

- Output is a set of nested **partitions**, starting with identity partition and ending with the complete partition
  - A "PARTITION" is a vector that associates each node with one and only one "group" (mutually exclusive)

- Different flavors based on how distance from a cluster to outside point/node is defined
  - Single linkage; connectedness; minimum
  - Complete linkage; diameter; maximum
  - Average, median, etc.

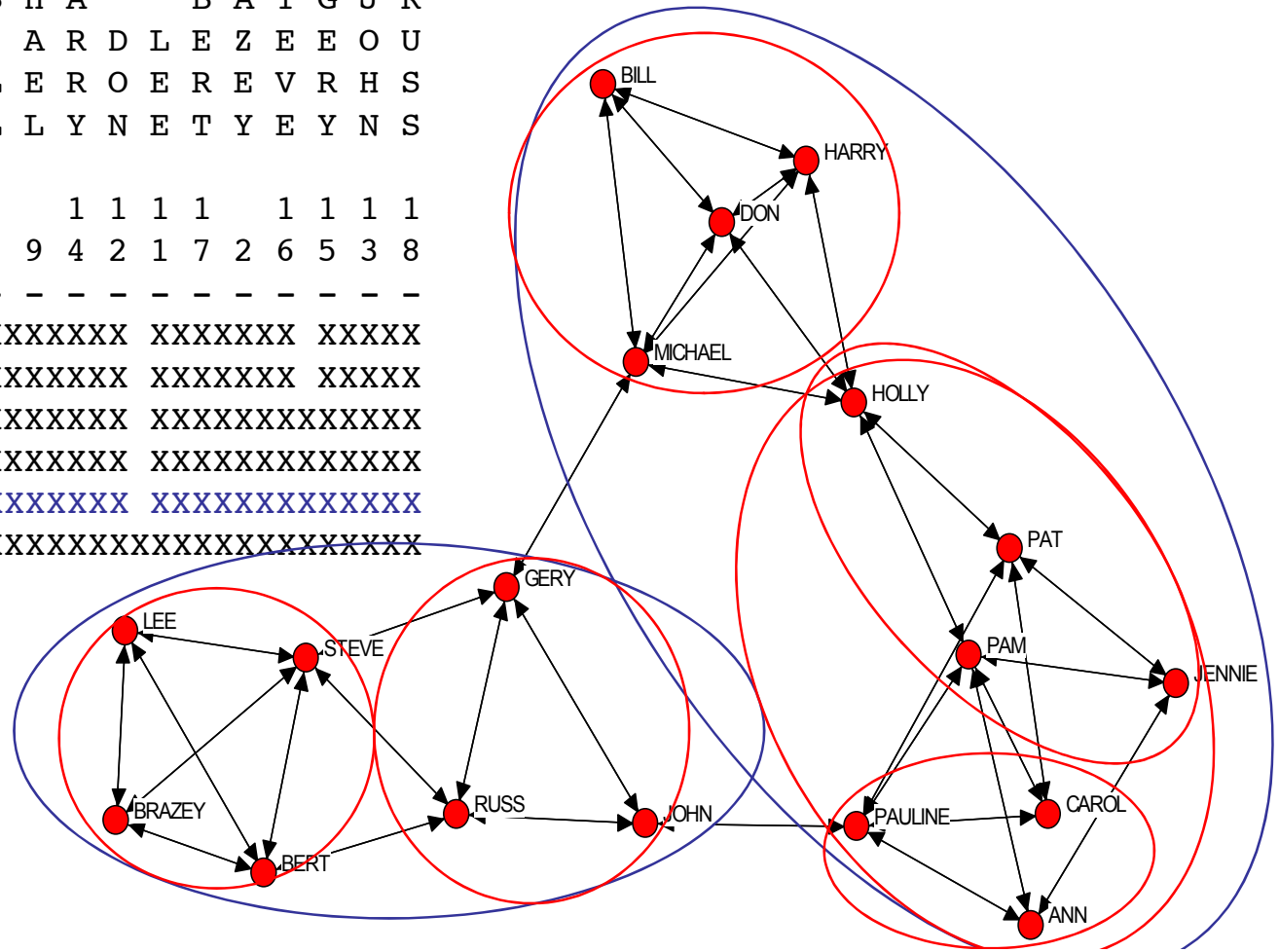
- BETTER: Compute geodesic distances first, then cluster the distance matrix (again using average method)
- Or cluster the structural equivalence matrix (tomorrow)

Geodesic Distances

		1	2	3	4	5	6	7	8	9	1 0	1 1	1 2	1 3	1 4	1 5	1 6	1 7	1 8
		H	B	C	P	P	J	P	A	M	B	L	D	J	H	G	S	B	R
1	HOLLY	0	4	2	1	1	2	2	2	1	2	4	1	3	1	2	3	4	3
2	BRAZEY	4	0	5	5	5	6	4	5	3	4	1	4	3	4	2	1	1	2
3	CAROL	2	5	0	1	1	2	1	2	3	4	5	3	2	3	3	4	4	3
4	PAM	1	5	1	0	2	1	1	1	2	3	5	2	2	2	3	4	4	3
5	PAT	1	5	1	2	0	1	1	2	2	3	5	2	2	2	3	4	4	3
6	JENNIE	2	6	2	1	1	0	2	1	3	4	6	3	3	3	4	5	5	4
7	PAULINE	2	4	1	1	1	2	0	1	3	4	4	3	1	3	2	3	3	2
8	ANN	2	5	2	1	2	1	1	0	3	4	5	3	2	3	3	4	4	3
9	MICHAEL	1	3	3	2	2	3	3	3	0	1	3	1	2	1	1	2	3	2
10	BILL	2	4	4	3	3	4	4	4	1	0	4	1	3	1	2	3	4	3
11	LEE	4	1	5	5	5	6	4	5	3	4	0	4	3	4	2	1	1	2
12	DON	1	4	3	2	2	3	3	3	1	1	4	0	3	1	2	3	4	3
13	JOHN	3	3	2	2	2	3	1	2	2	3	3	3	0	3	1	2	2	1
14	HARRY	1	4	3	2	2	3	3	3	1	1	4	1	3	0	2	3	4	3
15	GERY	2	2	3	3	3	4	2	3	1	2	2	2	1	2	0	1	2	1
16	STEVE	3	1	4	4	4	5	3	4	2	3	1	3	2	3	1	0	1	1
17	BERT	4	1	4	4	4	5	3	4	3	4	1	4	2	4	2	1	0	1
18	RUSS	3	2	3	3	3	4	2	3	2	3	2	3	1	3	1	1	1	0

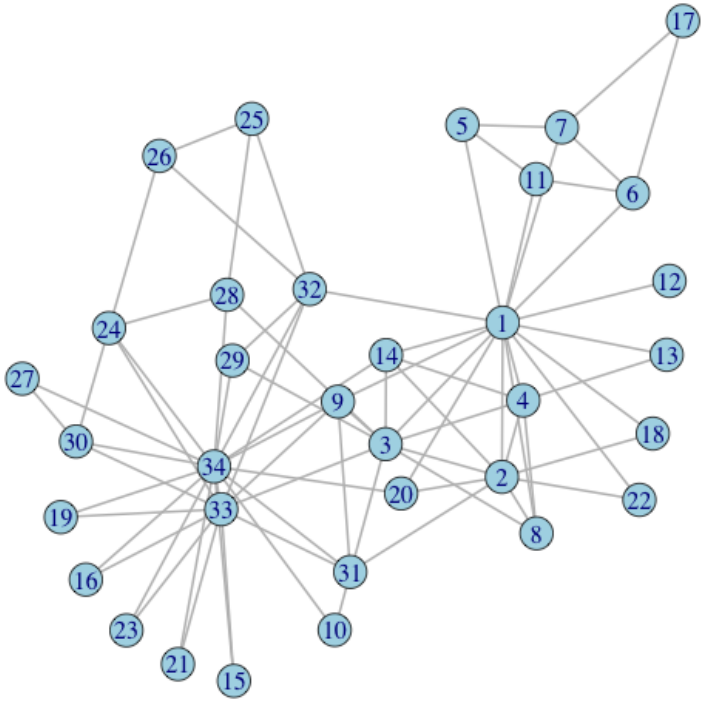
Clustering on Node Similarities | Agglomerative Hierarchical clustering

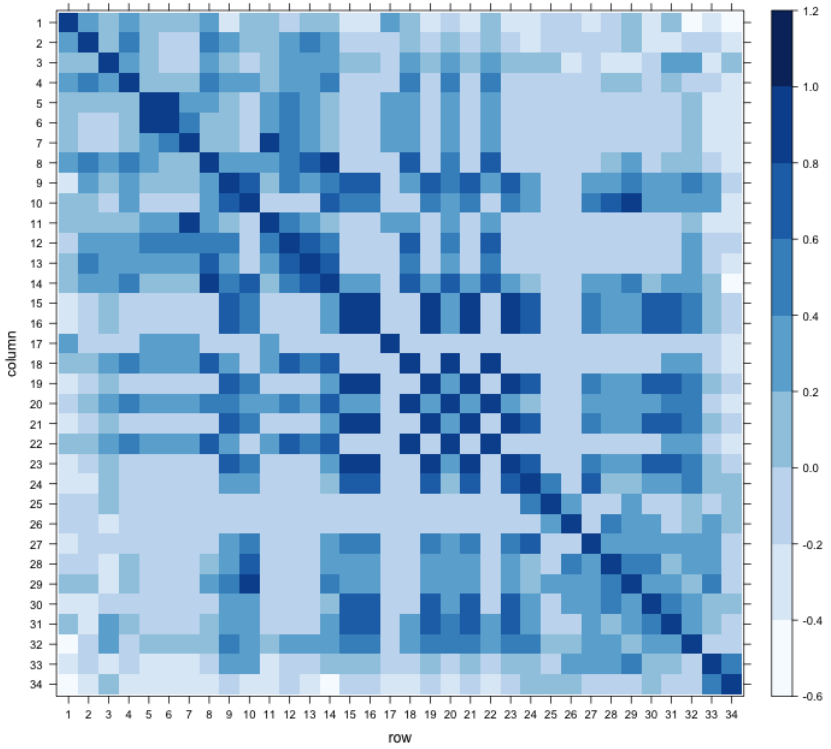
```
                P           M
                A     J     I           B
              C U   H E     C H       R S
              A L   O N   B H A     B A T G J R
          P R I P L N A I A R D L E Z E E O U
          A O N A L I N L E R O E R E V R H S
          T L E M Y E N L L Y N E T Y E Y N S

                          1   1 1 1 1   1 1 1 1
Level     5 3 7 4 1 6 8 0 9 4 2 1 7 2 6 5 3 8
-----     - - - - - - - - - - - - - - - - - - -
1.000     XXXXX XXX XXX XXXXXXX XXXXXXX XXXXX
1.333     XXXXX XXXXXXX XXXXXXX XXXXXXX XXXXX
1.457     XXXXX XXXXXXX XXXXXXX XXXXXXXXXXXXX
1.481     XXXXXXXXXXXXX XXXXXXX XXXXXXXXXXXXX
2.723     XXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXX
3.142     XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```



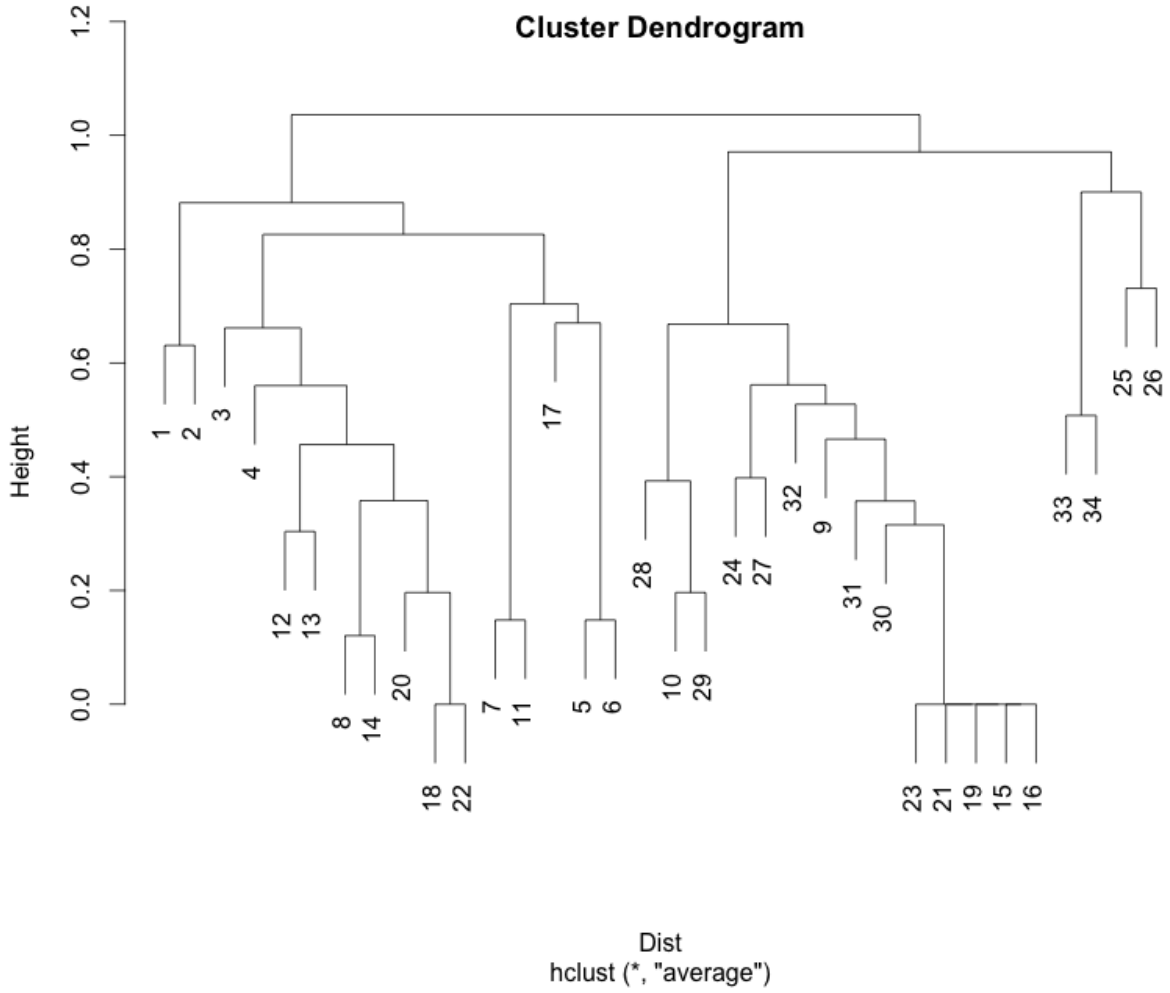
Clustering on Node Similarities | Agglomerative Hierarchical clustering

Zachary karate club



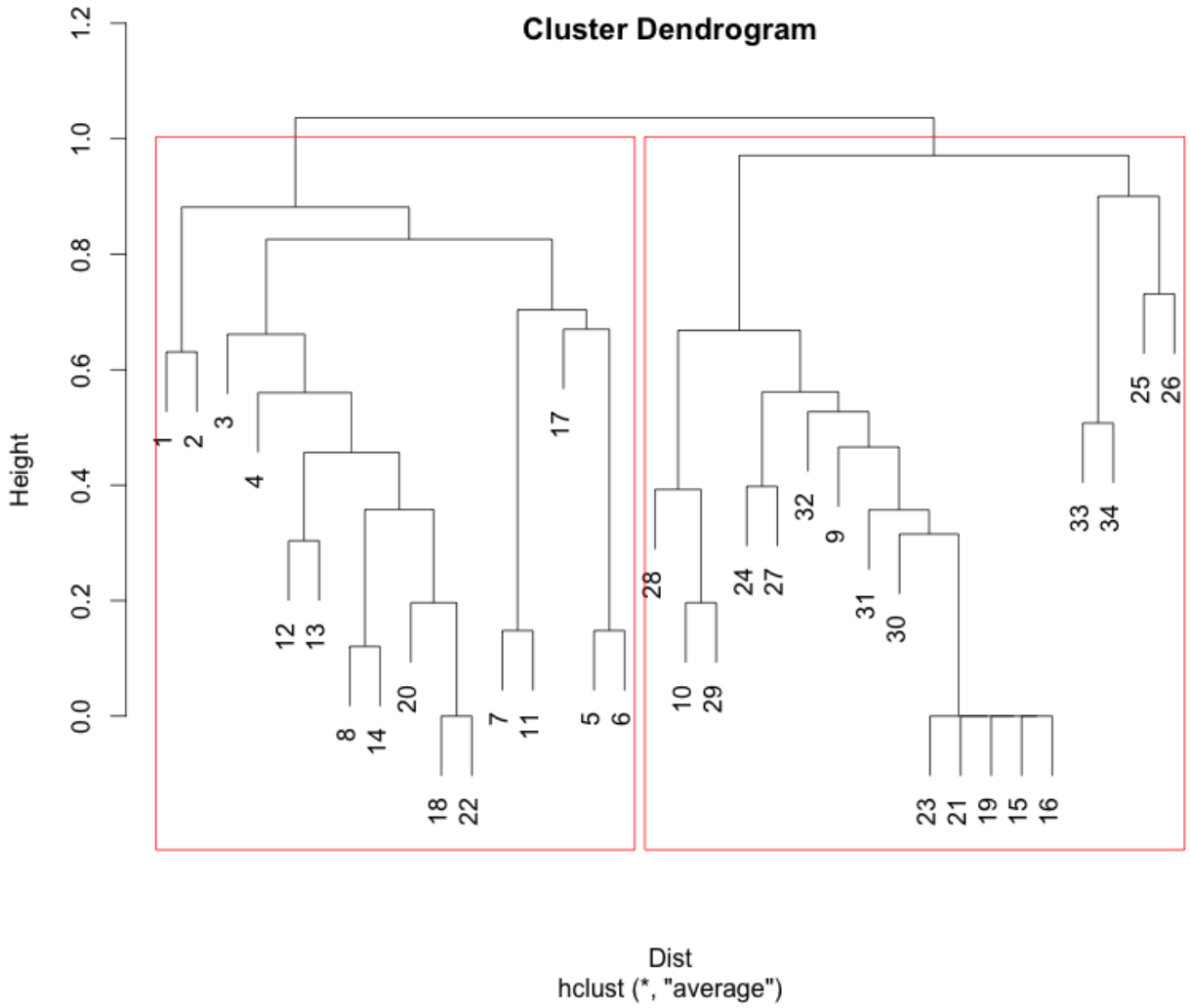


Clustering on Node Similarities | Agglomerative Hierarchical clustering

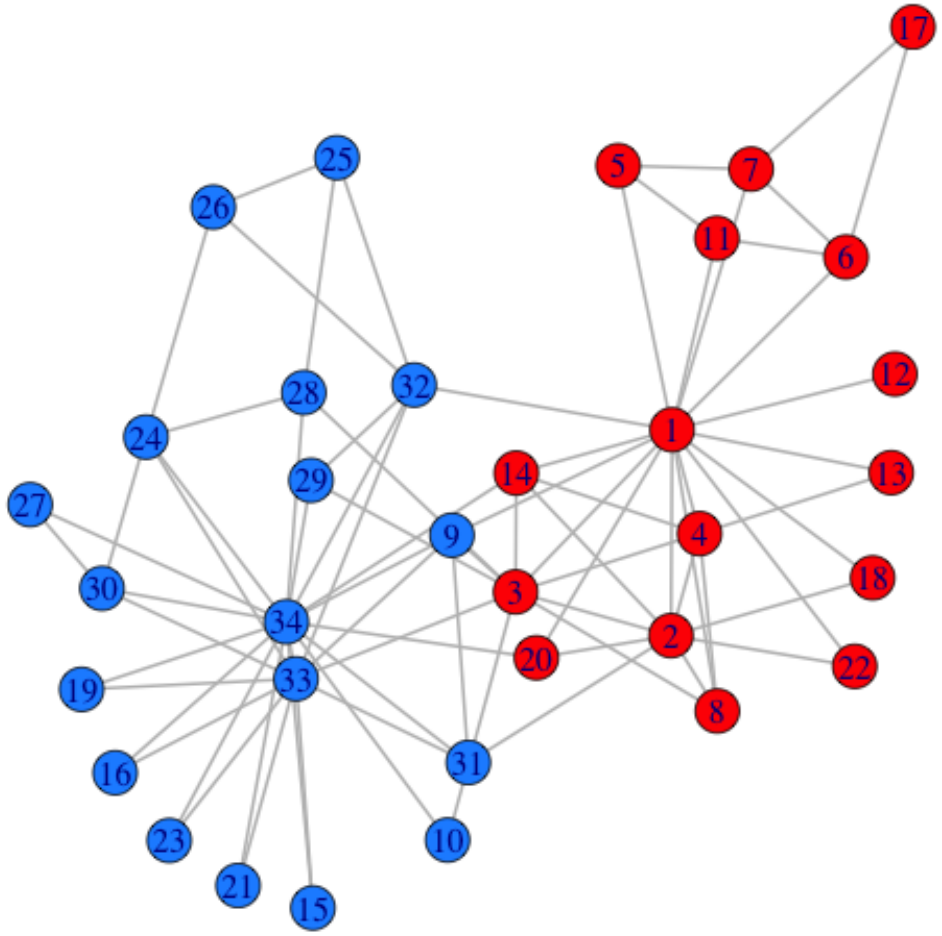


We can decide at what level we want to cut. Do we want very *fine* or very *coarse* communities?

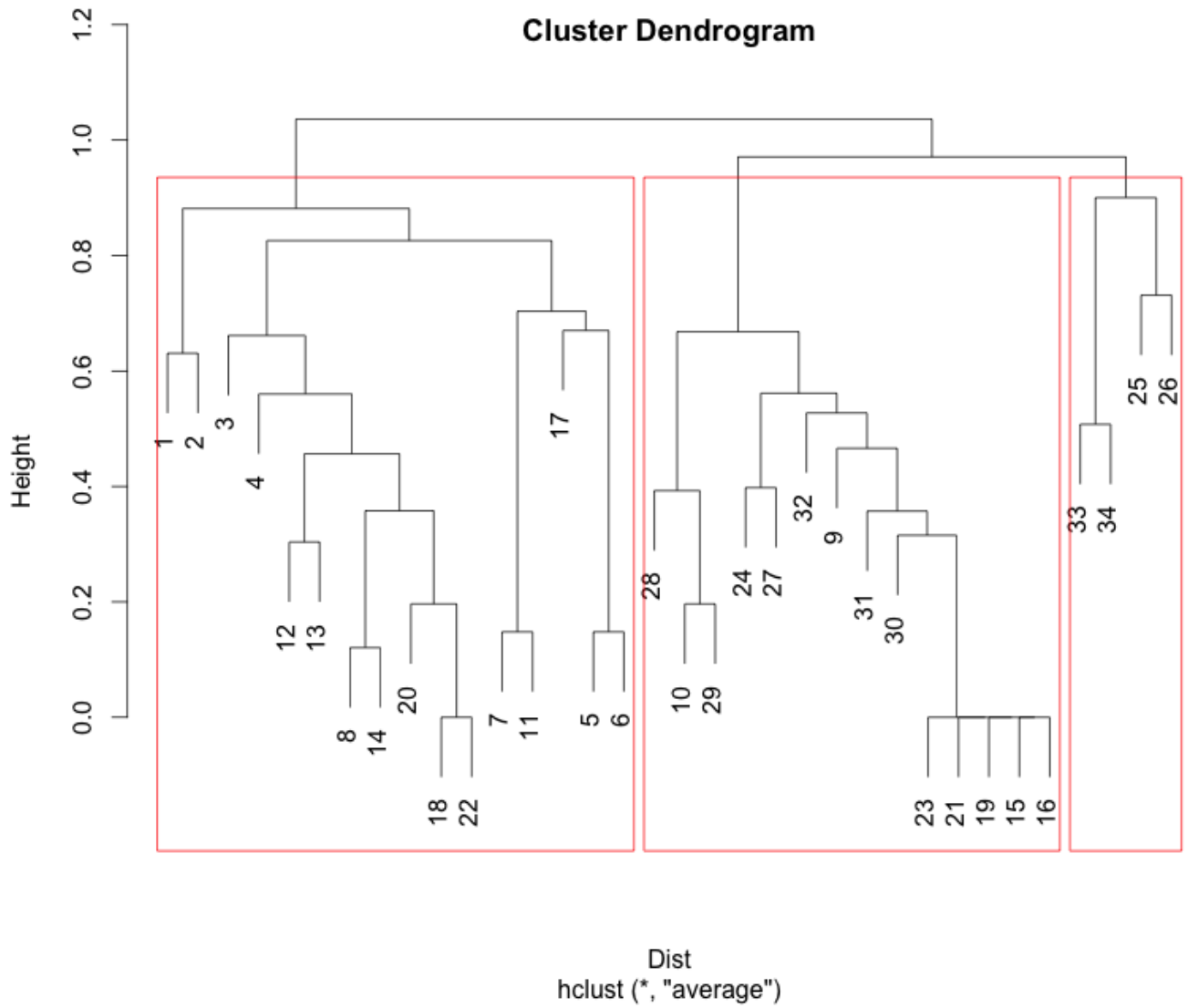
Clustering on Node Similarities | Agglomerative Hierarchical clustering



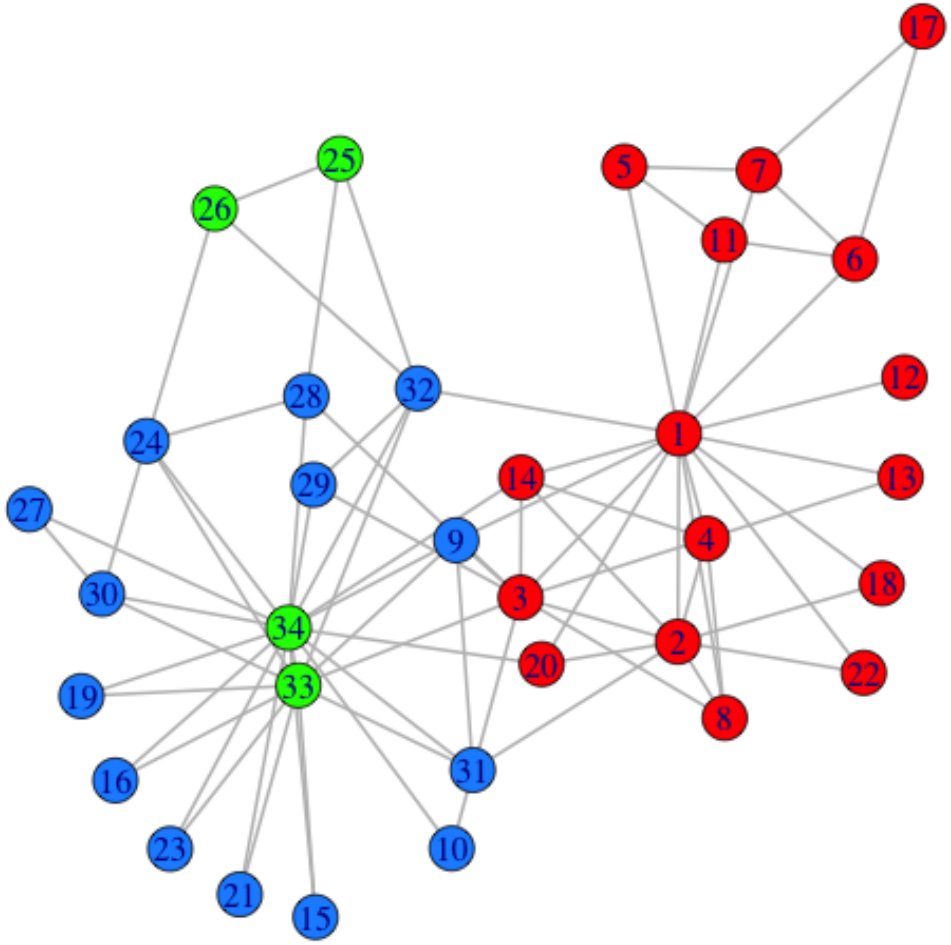
Clustering on Node Similarities | Agglomerative Hierarchical clustering



Clustering on Node Similarities | Agglomerative Hierarchical clustering



Clustering on Node Similarities | Agglomerative Hierarchical clustering

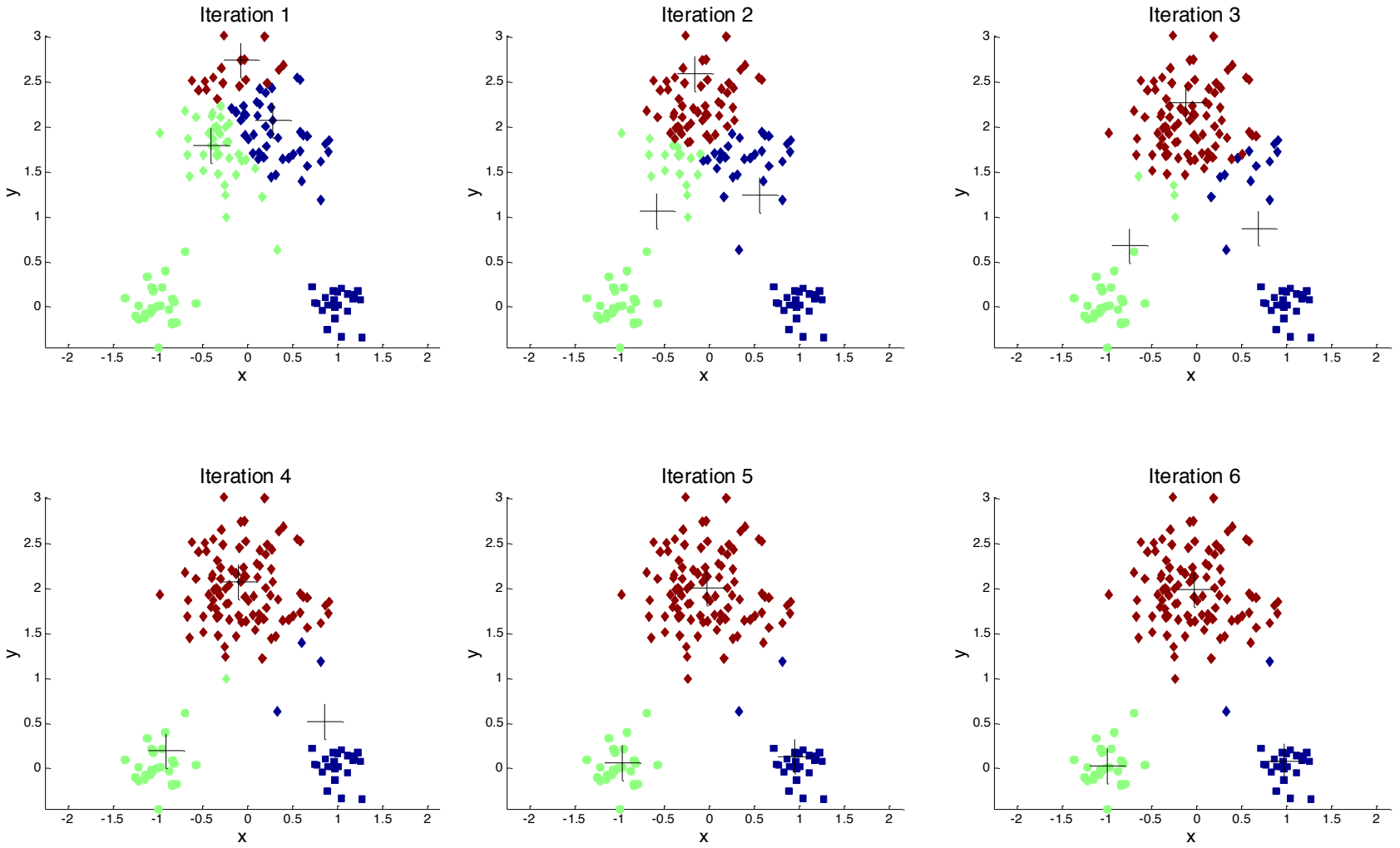


- K-means Clustering Algorithm
  - Each cluster is associated with a centroid (center point)
  - Each node is assigned to the cluster with the closest centroid

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Node Similarity| k-means clustering



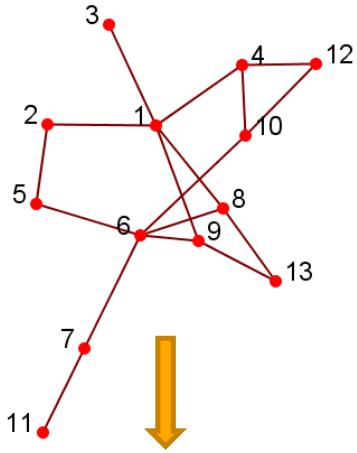
Node Similarity| Multidimensional Scaling

- **Latent-space models**: Transform the nodes in a network into a lower-dimensional space such that the distance or similarity between nodes are kept in the Euclidean space
- Multidimensional Scaling (MDS)
  - Given a network, construct a proximity matrix to denote the distance between nodes (e.g. geodesic distance)
  - Let D denotes the *square distance* between nodes
  - $S \in R^{n \times k}$ denotes the coordinates in the lower-dimensional space

$$SS^T = -\frac{1}{2}(I - \frac{1}{n}ee^T)D(I - \frac{1}{n}ee^T) = \Delta(D)$$

  - Objective: minimize the difference $\min \| \Delta(D) - SS^T \|_F$
  - Let $\Lambda = diag(\lambda_1, \cdots, \lambda_k)$ (the top-k eigenvalues of Δ), V the top-k eigenvectors
  - Solution: $S = V\Lambda^{1/2}$
- Apply k-means to S to obtain clusters

Node Similarity| Multidimensional Scaling



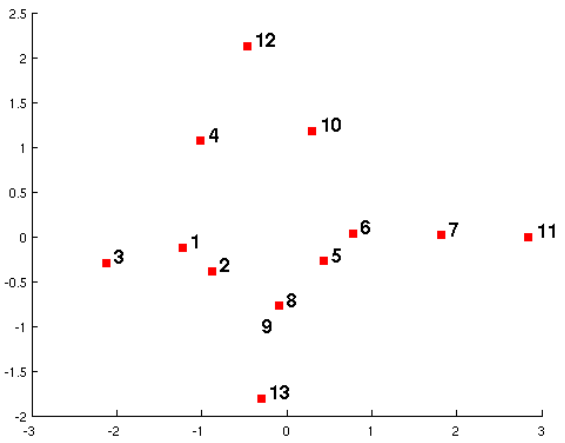
Geodesic Distance Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0	1	1	1	2	2	3	1	1	2	4	2	2
2	1	0	2	2	1	2	3	2	2	3	4	3	3
3	1	2	0	2	3	3	4	2	2	3	5	3	3
4	1	2	2	0	3	2	3	2	2	1	4	1	3
5	2	1	3	3	0	1	2	2	2	2	3	3	3
6	2	2	3	2	1	0	1	1	1	1	2	2	2
7	3	3	4	3	2	1	0	2	2	2	1	3	3
8	1	2	2	2	2	1	2	0	2	2	3	3	1
9	1	2	2	2	2	1	2	2	0	2	3	3	1
10	2	3	3	1	2	1	2	2	2	0	3	1	3
11	4	4	5	4	3	2	1	3	3	3	0	4	4
12	2	3	3	1	3	2	3	3	3	1	4	0	4
13	2	3	3	3	3	2	3	1	1	3	4	4	0

MDS

S

-1.22	-0.12
-0.88	-0.39
-2.12	-0.29
-1.01	1.07
0.43	-0.28
0.78	0.04
1.81	0.02
-0.09	-0.77
-0.09	-0.77
0.30	1.18
2.85	0.00
-0.47	2.13
-0.29	-1.81

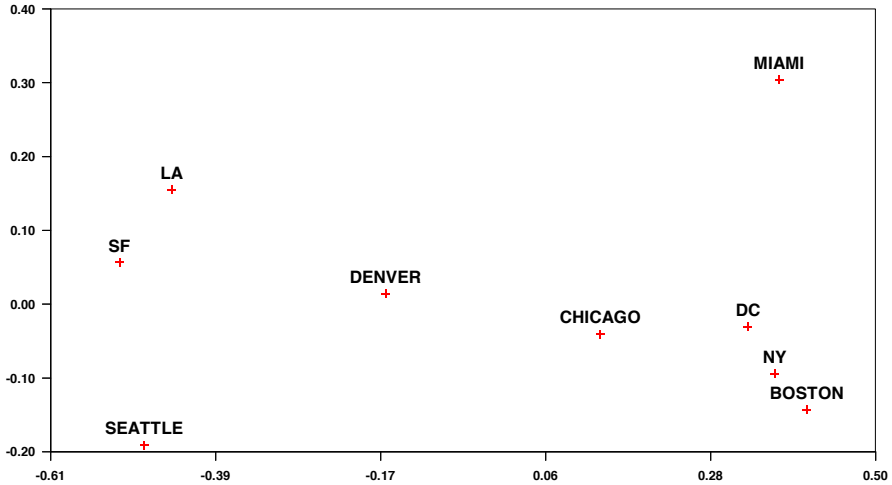


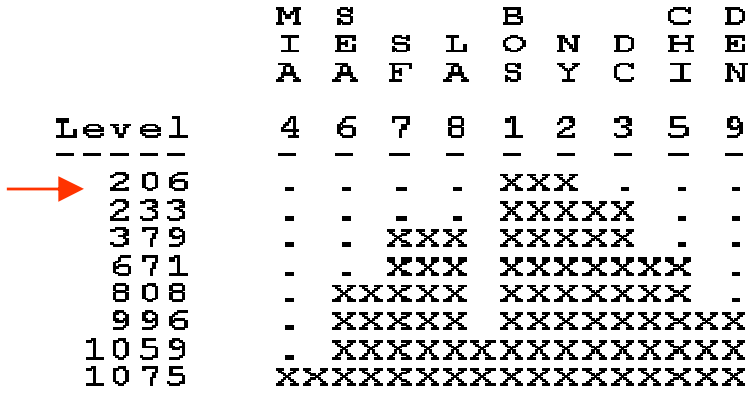
k-means

1, 2, 3, 4, 10, 12

5, 6, 7, 8, 9, 11, 13

Node Similarity| Multidimensional Scaling

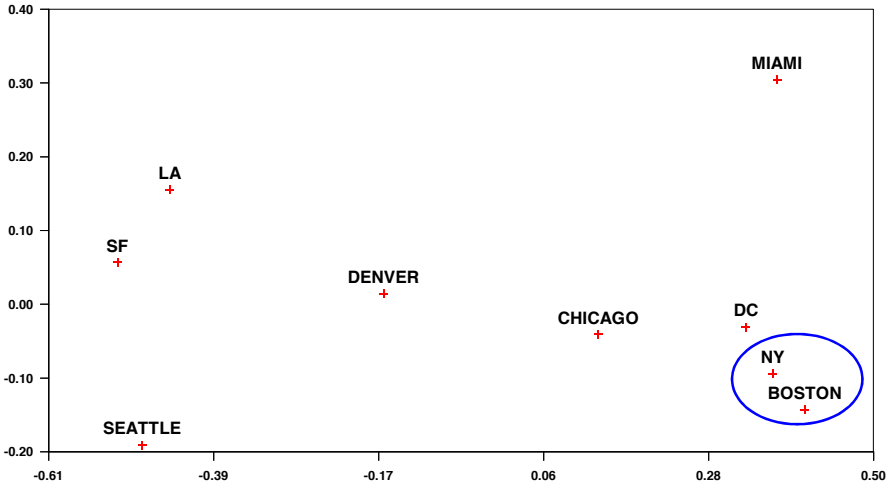


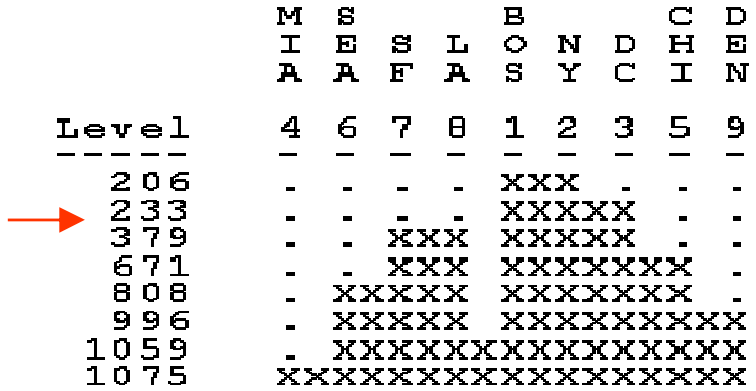


	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

Closest distance is NY-BOS = 206, so merge these.

Node Similarity| Multidimensional Scaling

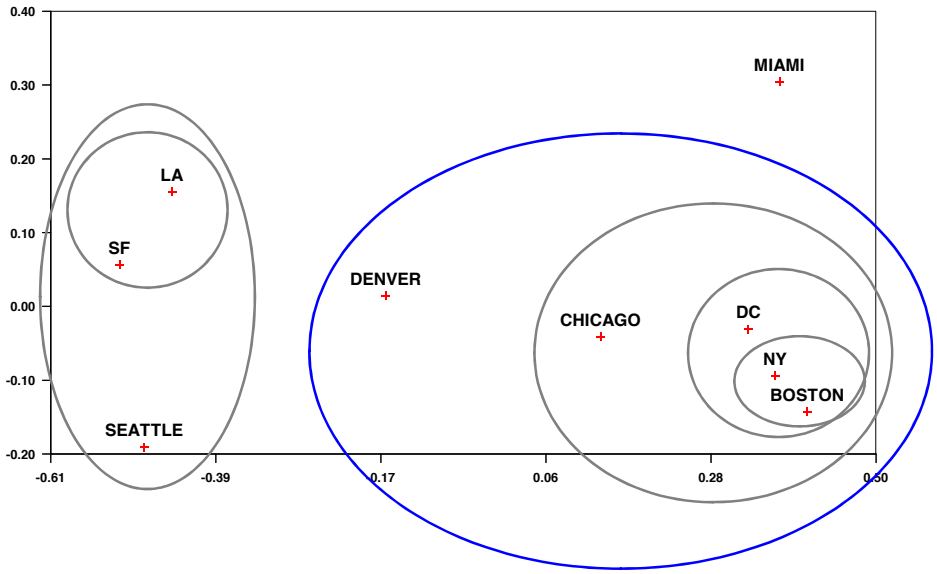


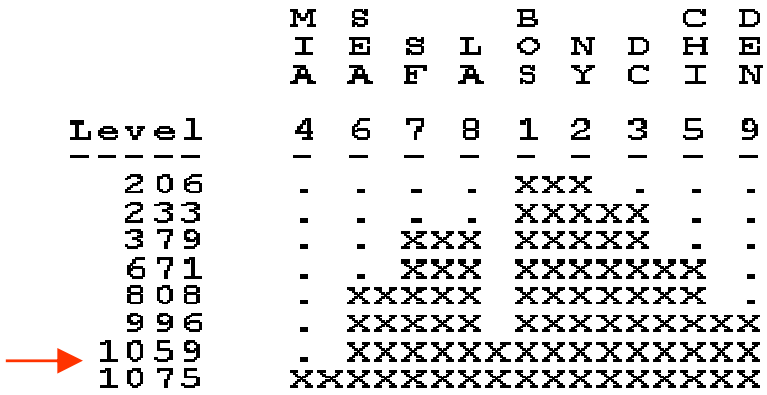


	BOS NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/ NY	0	233	1308	802	2815	2934	2786	1771
DC	233	0	1075	671	2684	2799	2631	1616
MIA	1308	1075	0	1329	3273	3053	2687	2037
CHI	802	671	1329	0	2013	2142	2054	996
SEA	2815	2684	3273	2013	0	808	1131	1307
SF	2934	2799	3053	2142	808	0	379	1235
LA	2786	2631	2687	2054	1131	379	0	1059
DEN	1771	1616	2037	996	1307	1235	1059	0

Closest pair is DC to BOSNY combo @ 233. So merge these.

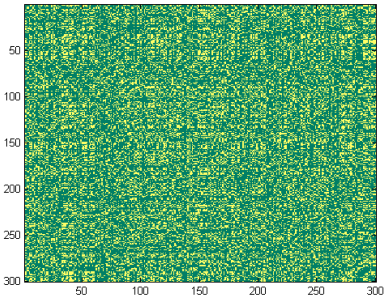
Node Similarity| Multidimensional Scaling



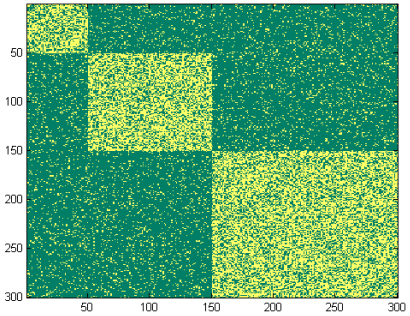


	BOS/ NY/D C/CHI /DEN	MIA	SF/LA /SEA
BOS/NY/DC/ CHI/DEN	0	1075	1059
MIA	1075	0	2687
SF/LA/SEA	1059	2687	0

Node Similarity| Block-Model Approximation



After Reordering



Network Interaction Matrix Block Structure

- Objective: Minimize the difference between an interaction matrix and a block structure

$$\min_{S,\Sigma} \|A - S\Sigma S^T\|_F$$

$$s.t. \quad S \in \{0,1\}^{n\times k}, \Sigma \in R^{k\times k} \text{ is diagonal}$$

S is a community indicator matrix

- Challenge: S is discrete, difficult to solve
- Relaxation: Allow S to be continuous satisfying $S^T S = I_k$
- Solution: the top eigenvectors of A
- Post-Processing: Apply k-means to S to find the partition

Hierarchy-Centric | Community Detection **Divisive Algorithms**

Hierarchy-Centric | Community Detection **Divisive Algorithms**

Goal is to build a hierarchical structure of communities based on network topology.

This now becomes a **graph partitioning** problem:

- we now focus on the edges rather than on similarity of the nodes;
- we want to cut as few edges as possible to see the graph split and fall apart into the groups of nodes that compose it.
- graph partitioning is NP-hard (Nondeterministic Polynomial time) – a class to classify complexity of problems.

e.g. (p) can you sort these cubes by color? sure, easy.

(np-hard) solve this sudoku puzzle; okay; after a long time, it's solved.

(np) can you check if the solution for the sudoku puzzle is valid/correct? yes, easy.

- Number of all possible partitions of a graph (n-th Bell number)

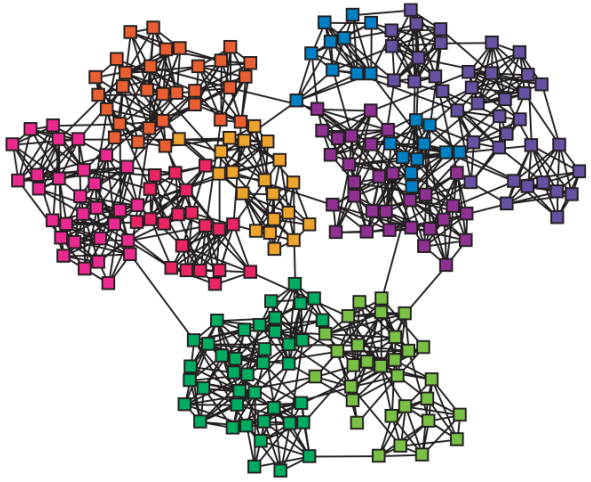
$$B_n = \sum_{k=1}^{n} S(n,k)$$

$$B_{20} = 5,832,742,205,057$$

Hierarchy-Centric | Heuristic Approach

Focus on edges that connect communities.
Edge betweenness -number of shortest paths $\sigma_{st}(e)$ going through edge e

$$C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$$



Newman-Girvan, 2004

Algorithm: Edge Betweenness

Input: graph G(V,E)

Output: Dendrogram

repeat

  For all $e \in E$ compute edge betweenness $C_B(e)$;

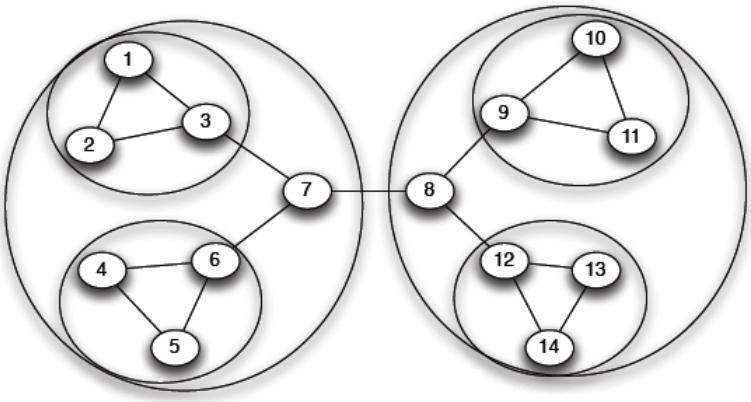
  remove edge e_i with largest $C_B(e_i)$;

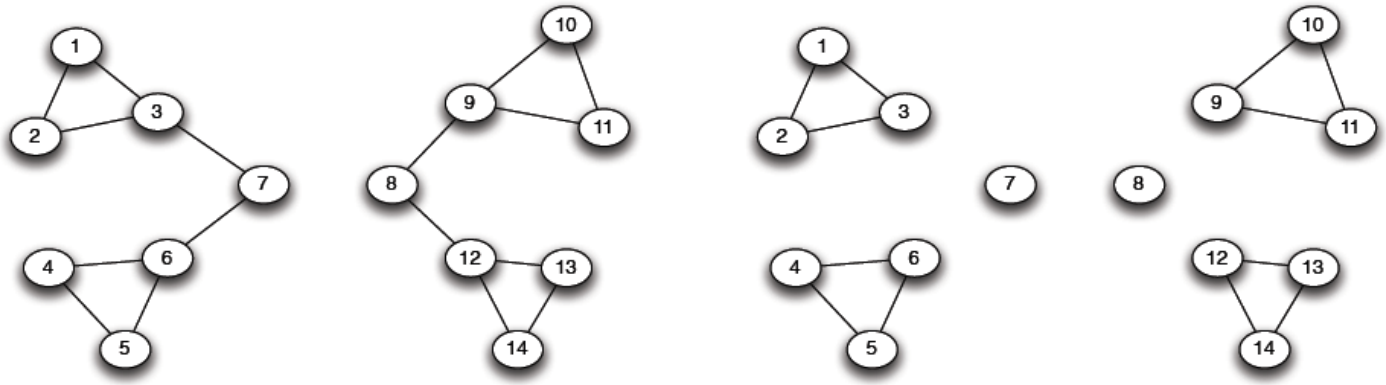
until *edges left*;

Construct communities by progressively removing edges

Hierarchy-Centric |**Girvan-Newman** Edge Betweenness algorithm

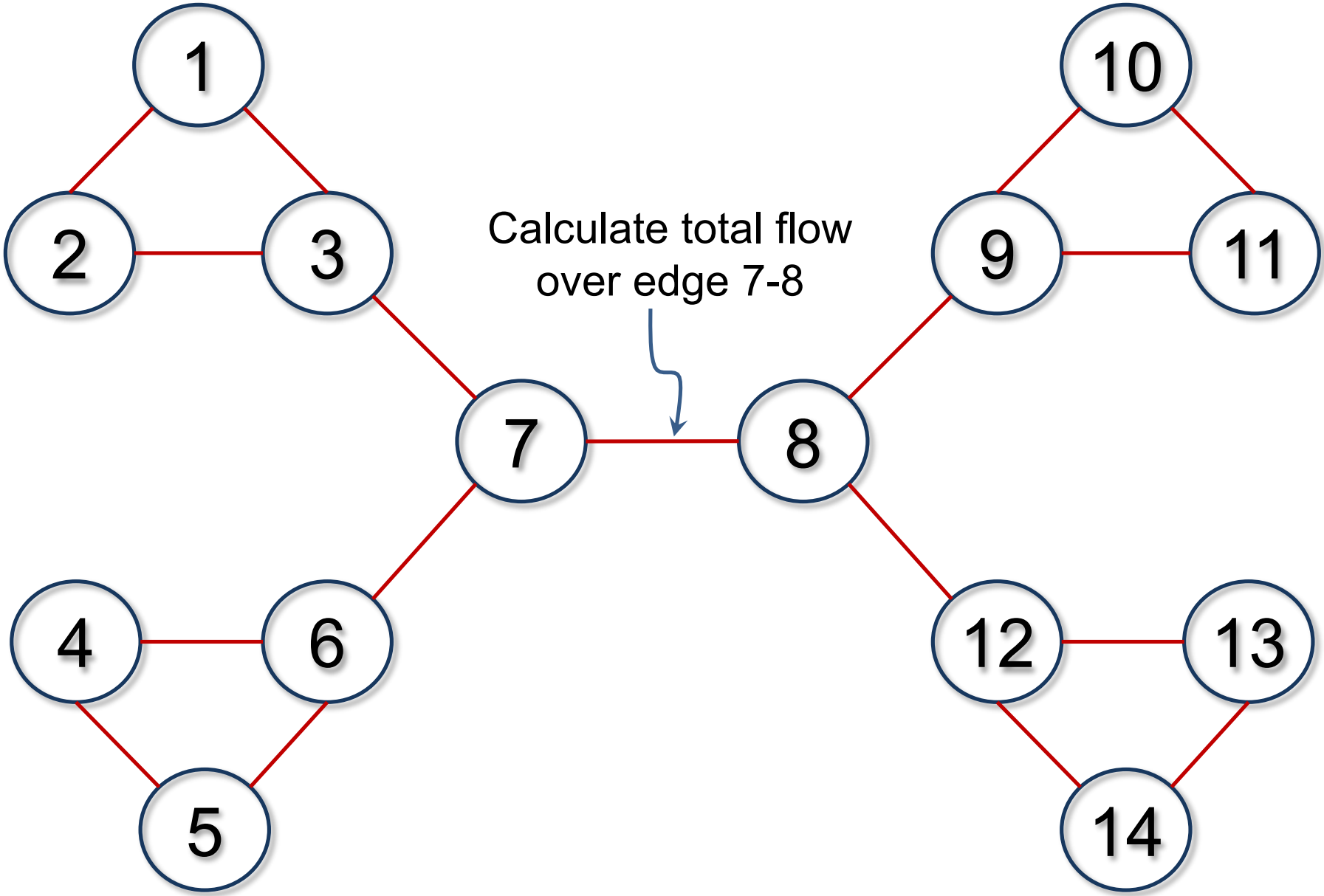
- successively remove edges of highest betweenness (the bridges, or local bridges), breaking up the network into separate components

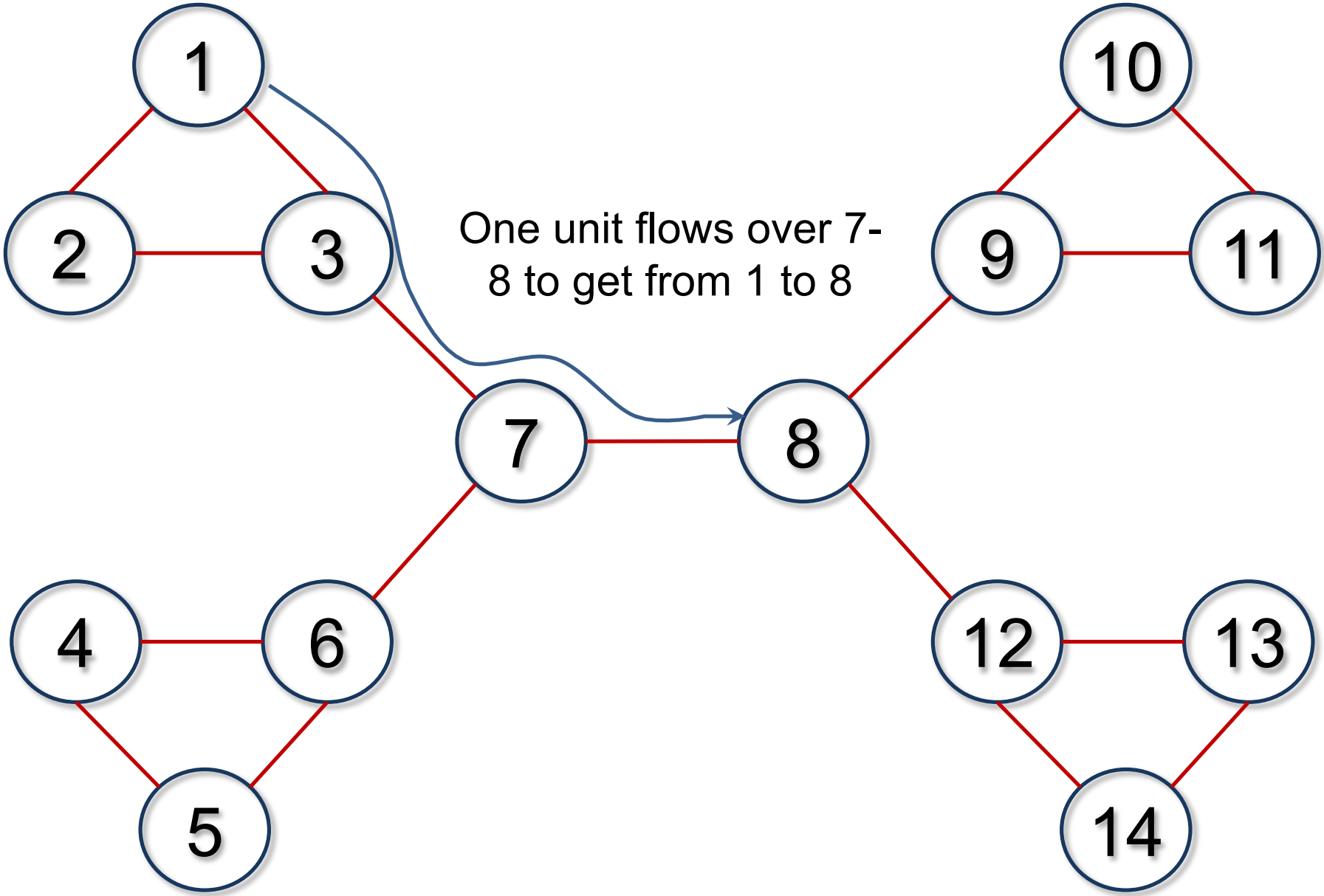


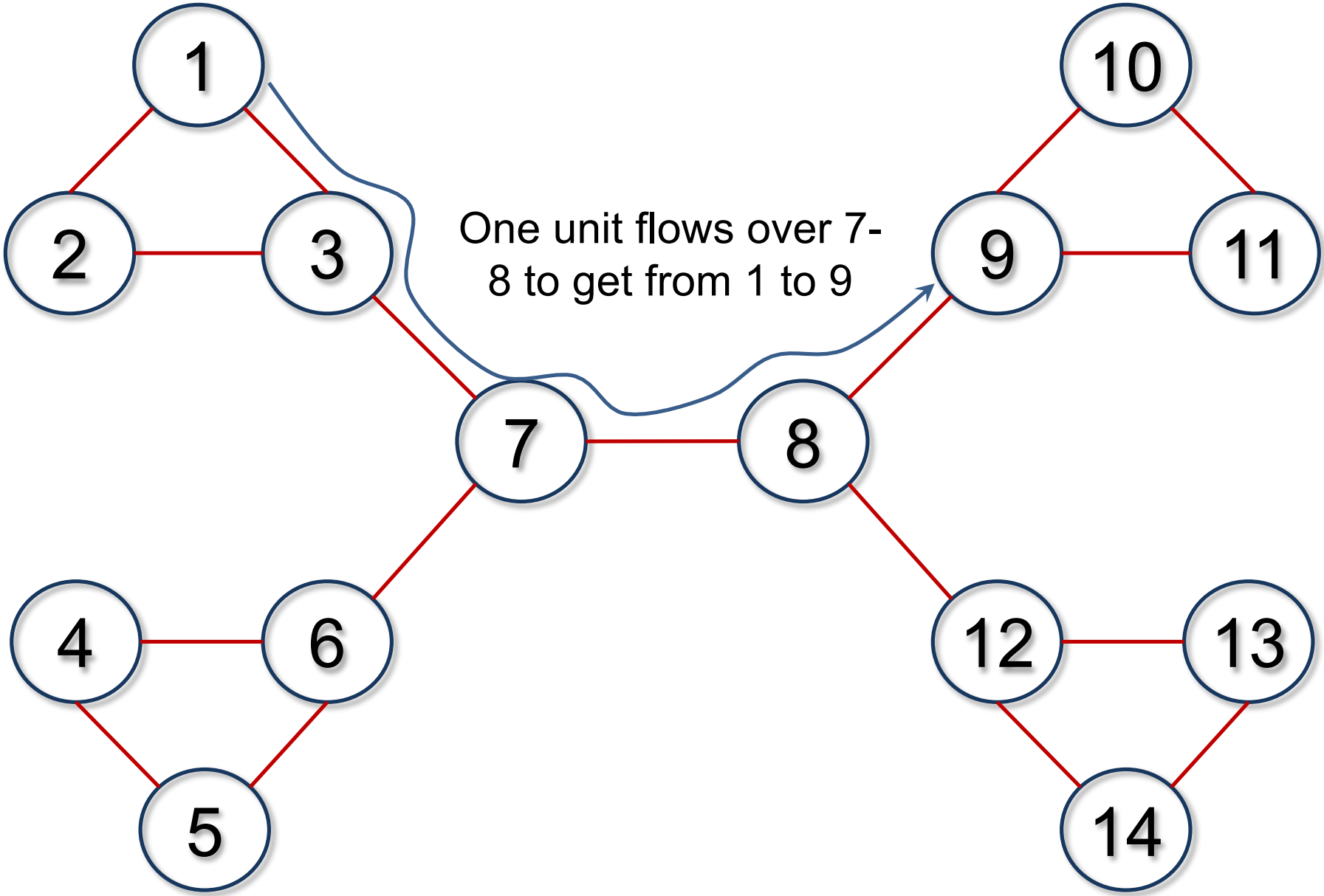


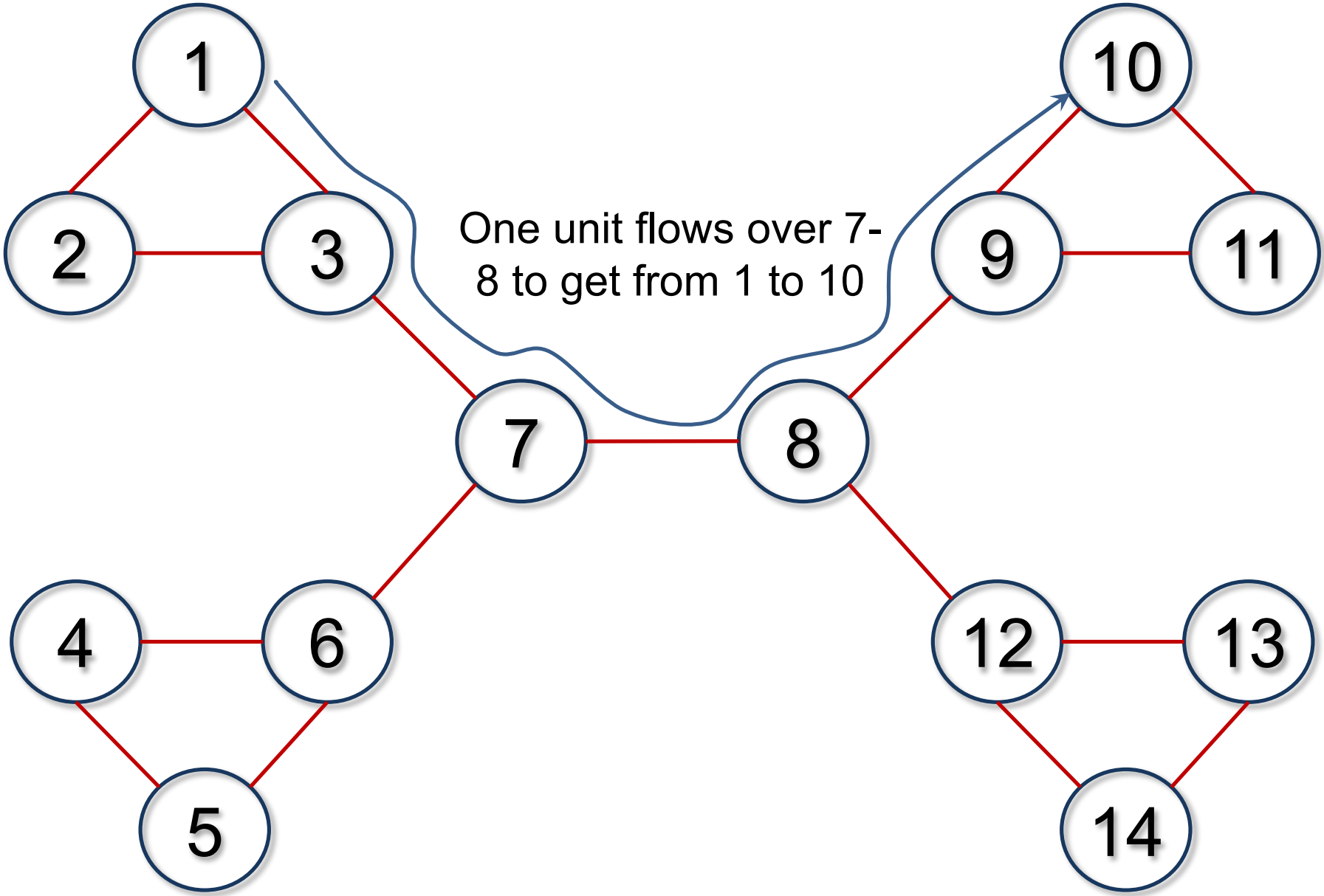
(a) *Step 1* (b) *Step 2*

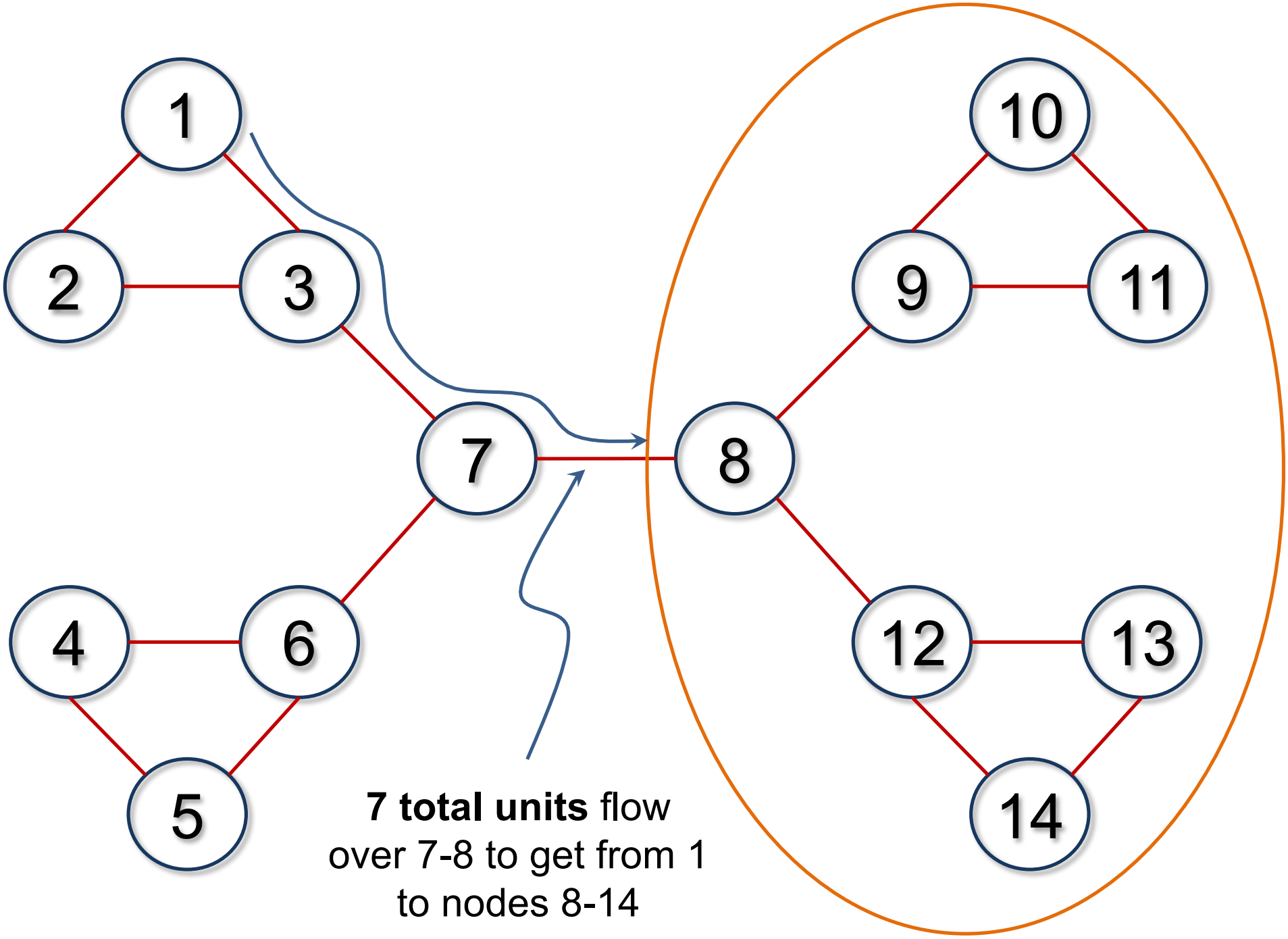
how do we calculate edge betweenness?

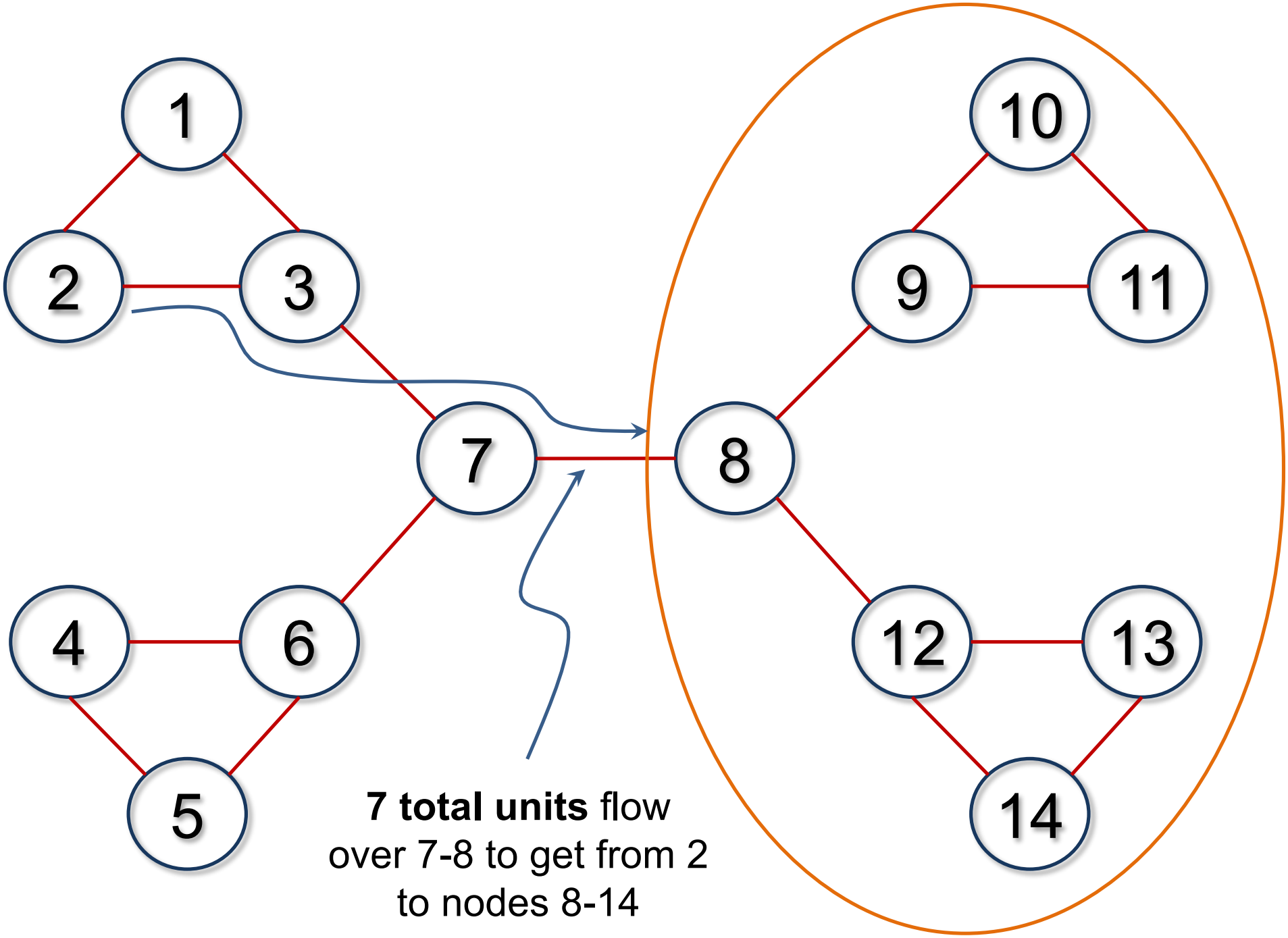


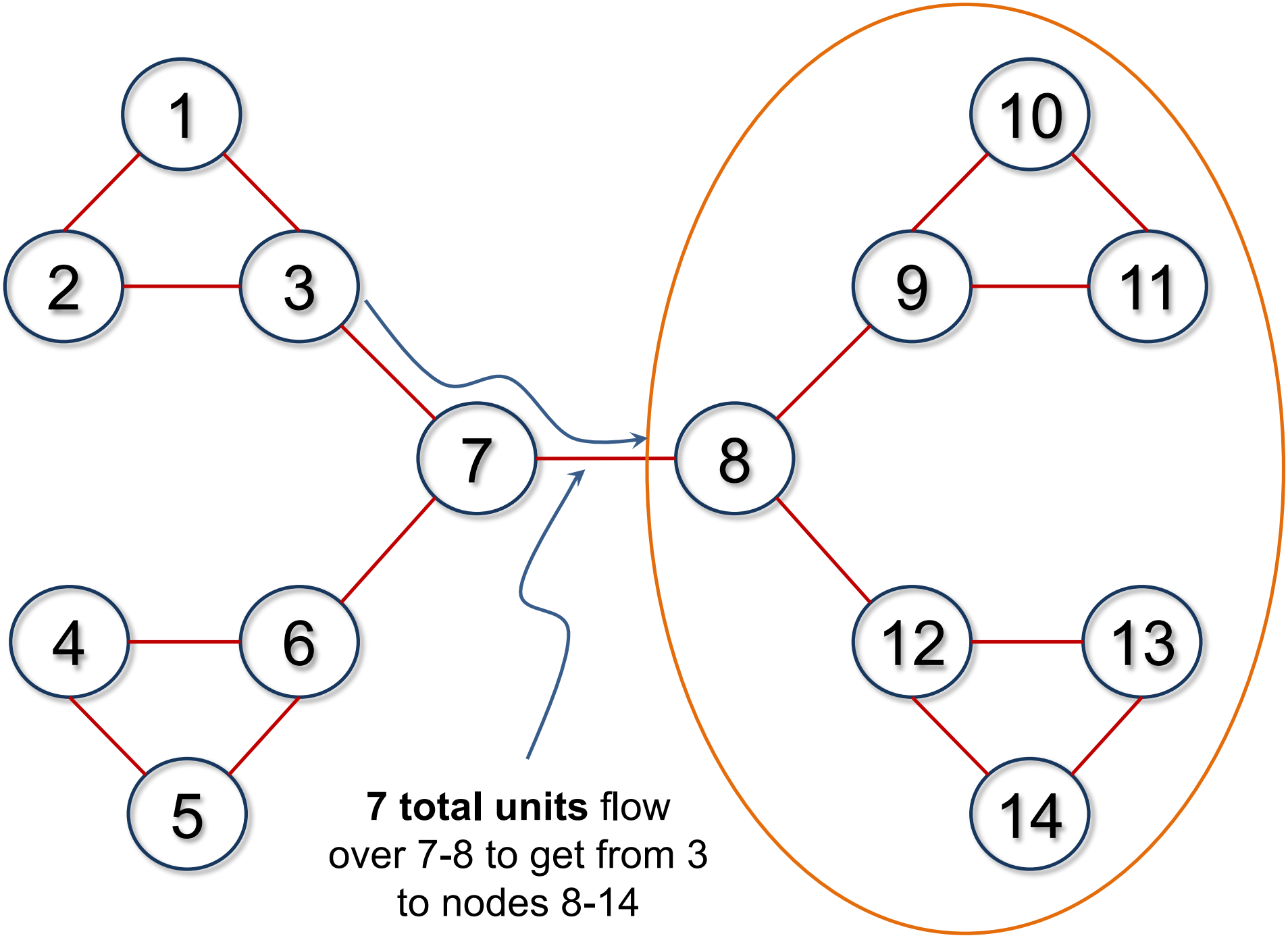

1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8 to get from 1 to 8

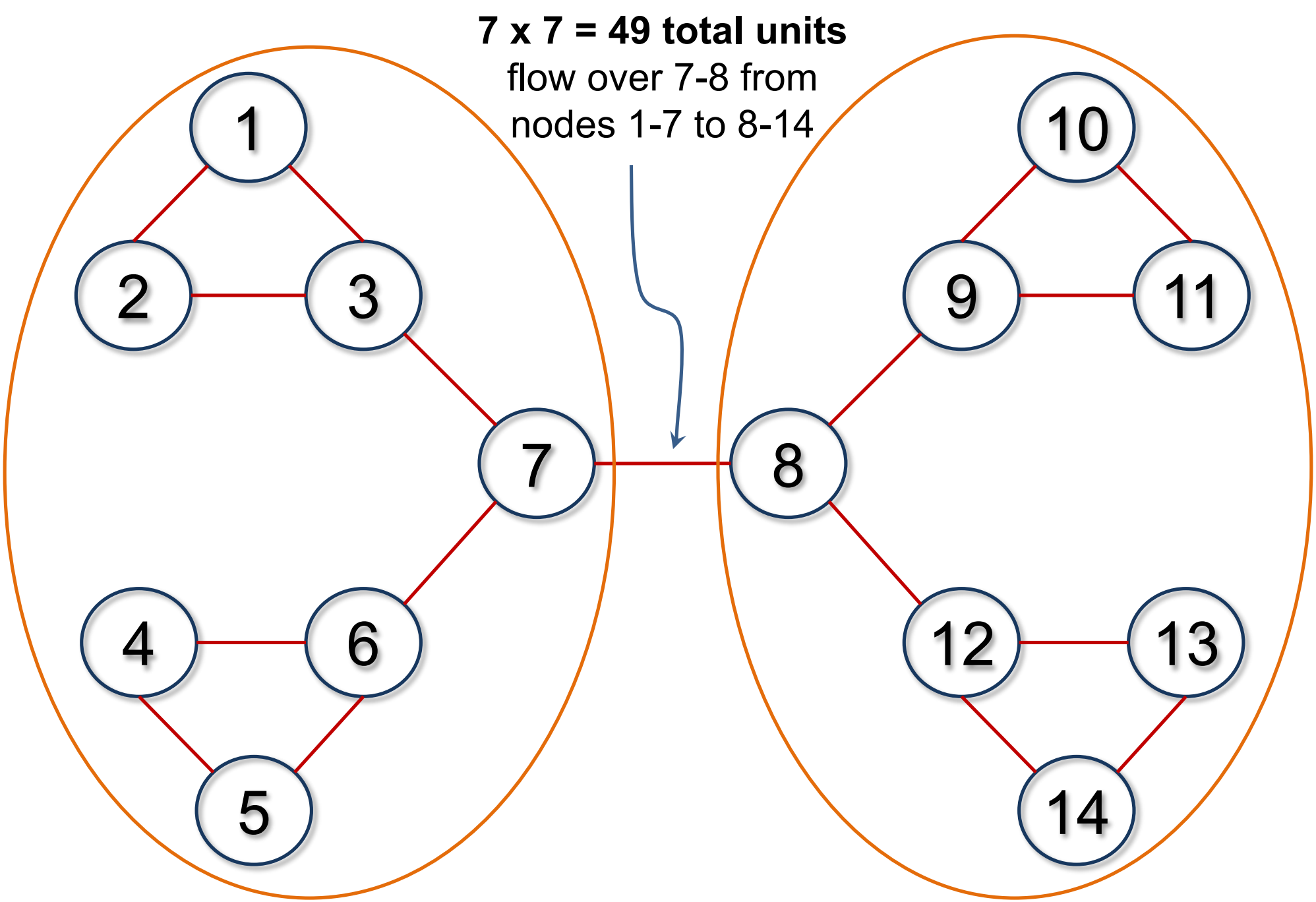

1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8 to get from 1 to 9

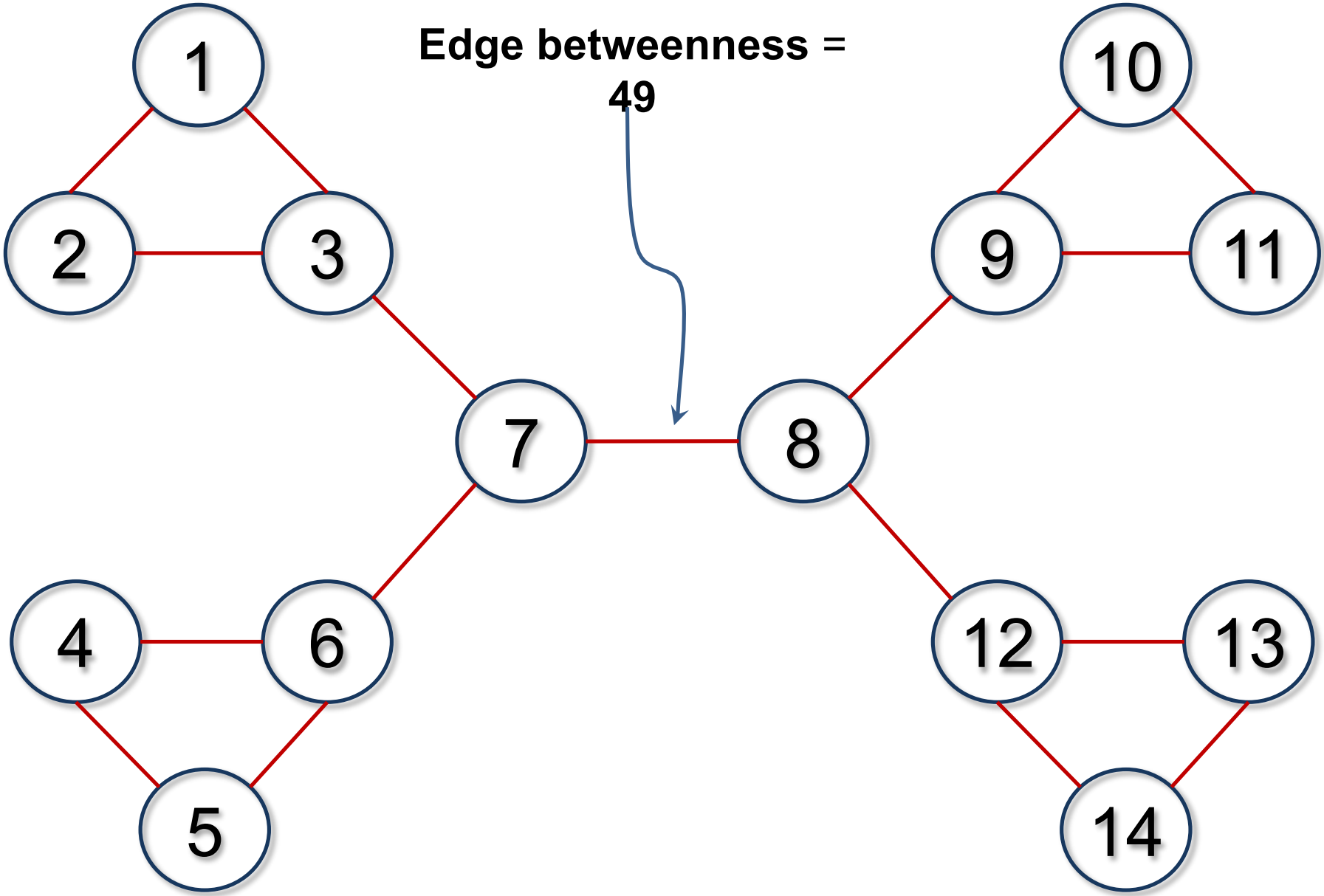

1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8 to get from 1 to 10

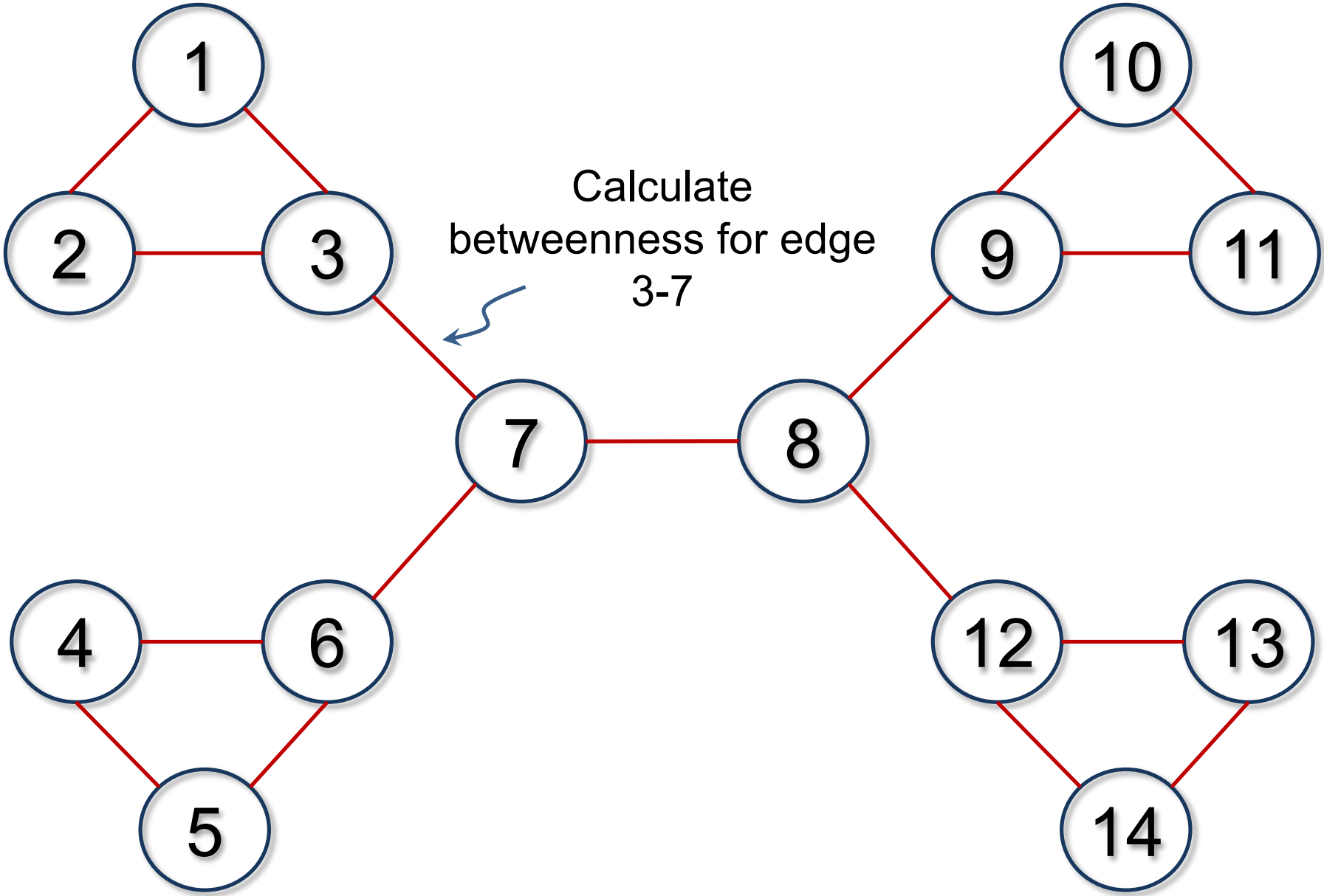

1
2
3
4
5
6
7
8
9
10
11
12
13
14
7 total units flow
over 7-8 to get from 1
to nodes 8-14

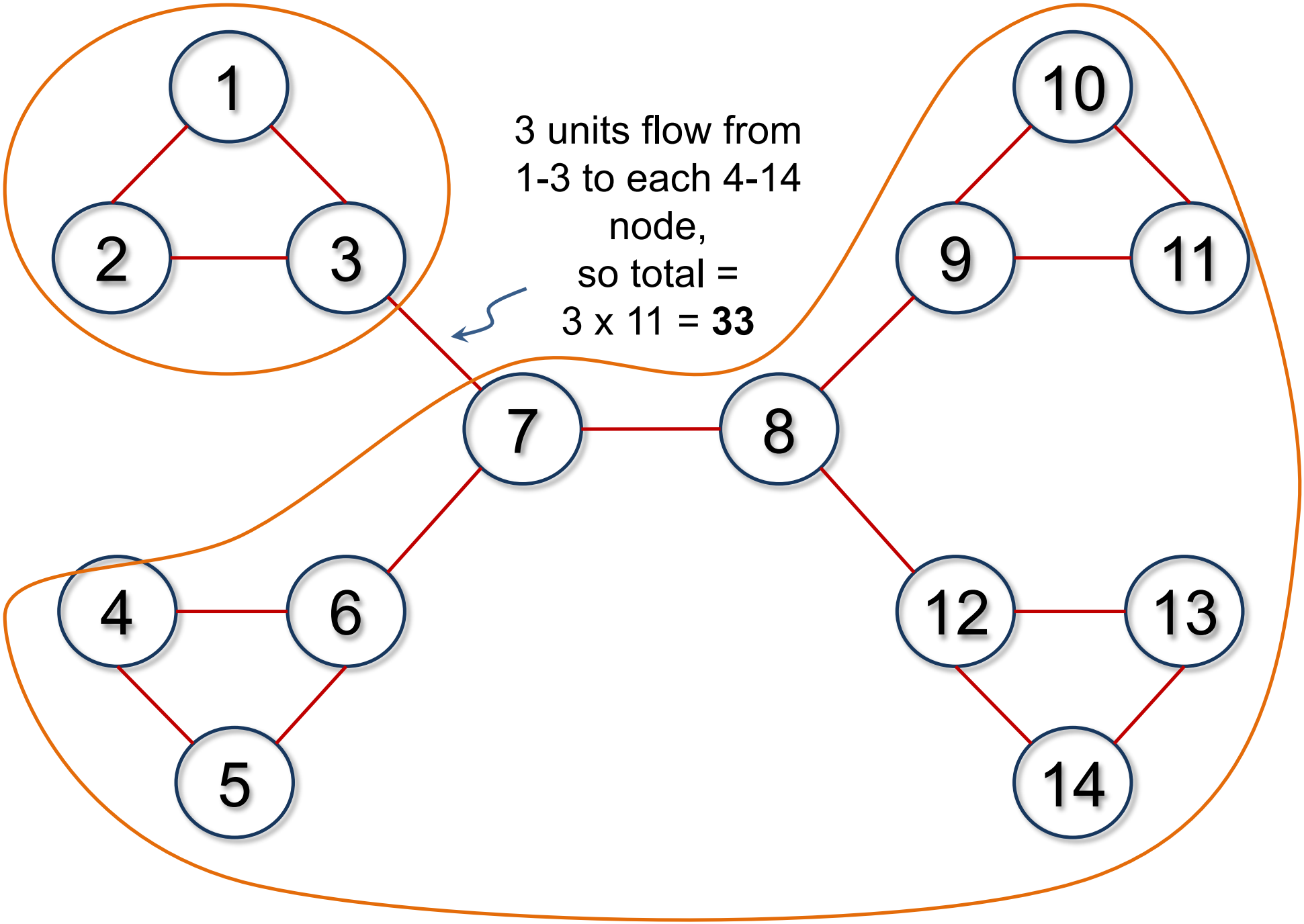

1
2
3
4
5
6
7
8
9
10
11
12
13
14
7 total units flow
over 7-8 to get from 2
to nodes 8-14

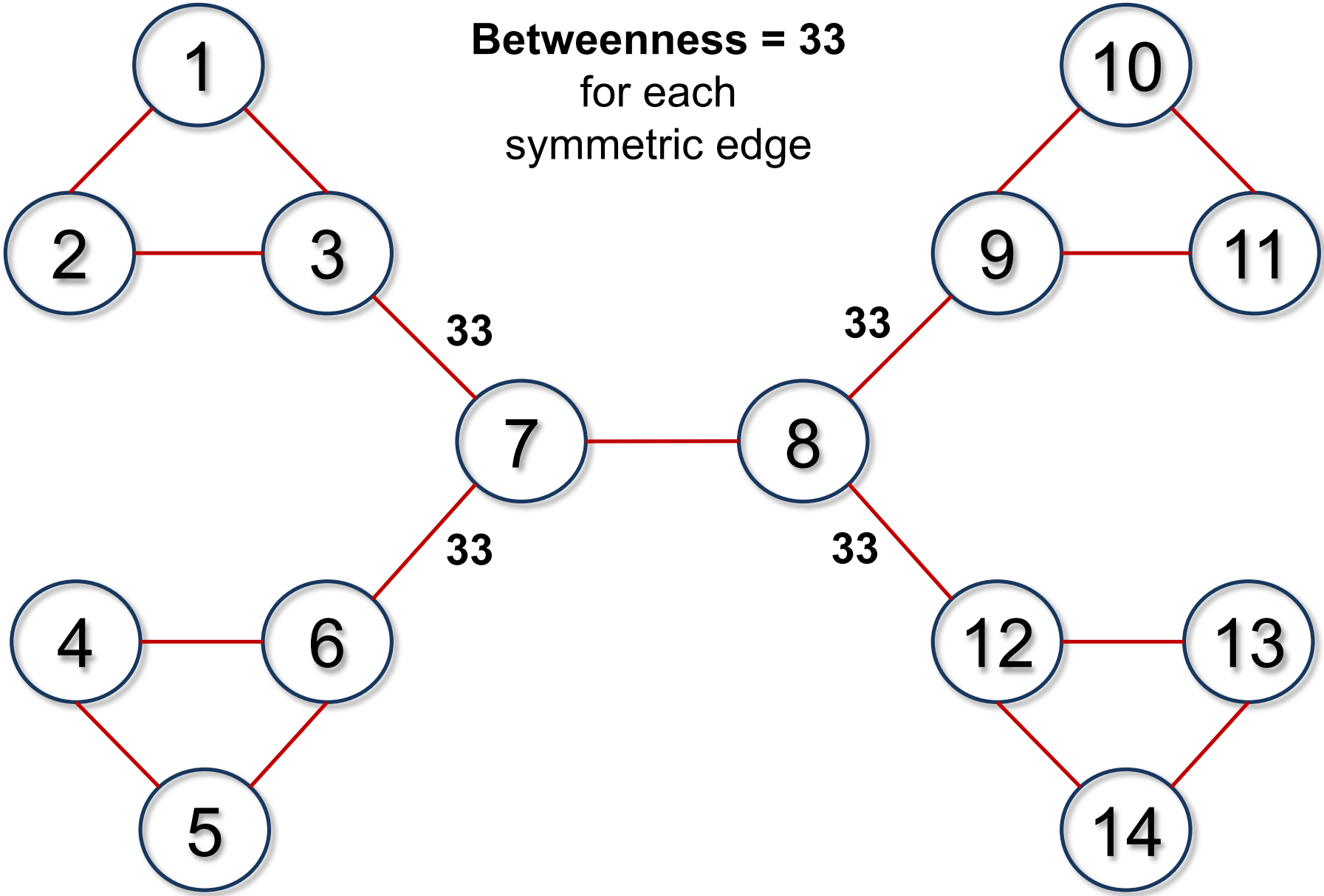

1
2
3
4
5
6
7
8
9
10
11
12
13
14
7 total units flow
over 7-8 to get from 3
to nodes 8-14

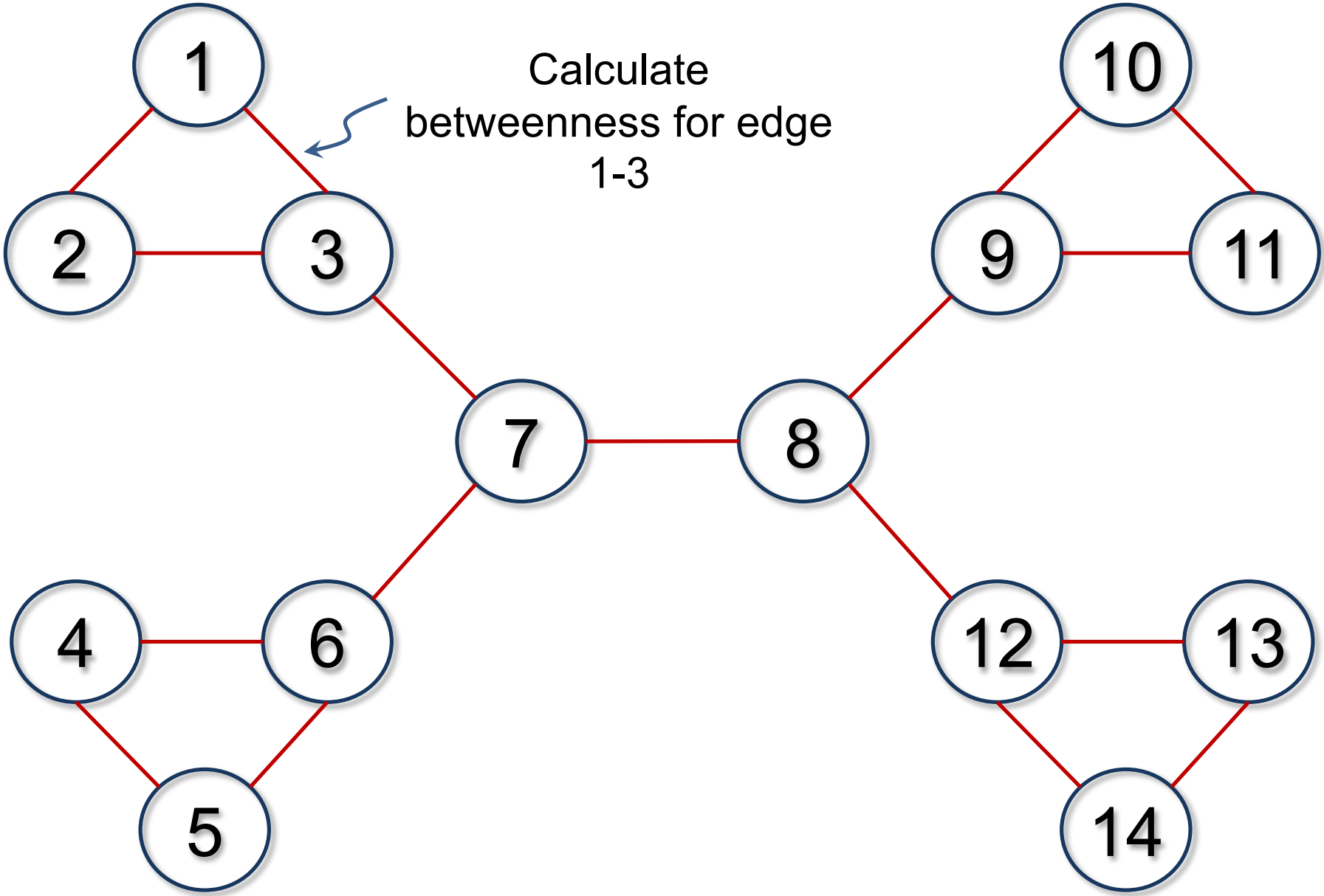

7 x 7 = 49 total units
flow over 7-8 from
nodes 1-7 to 8-14
1
2
3
4
5
6
7
8
9
10
11
12
13
14

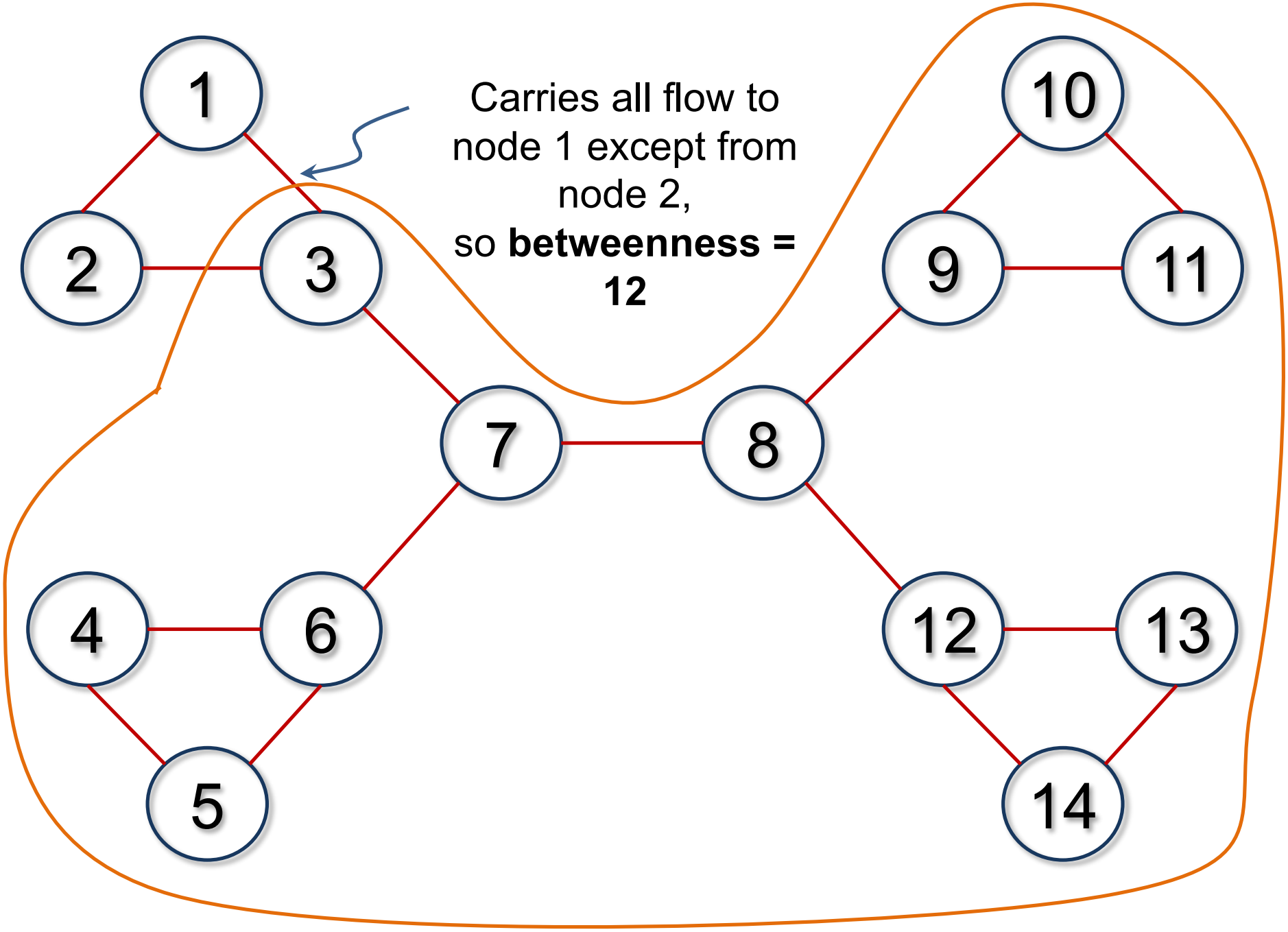

Edge betweenness =
49
1
2
3
4
5
6
7
8
9
10
11
12
13
14

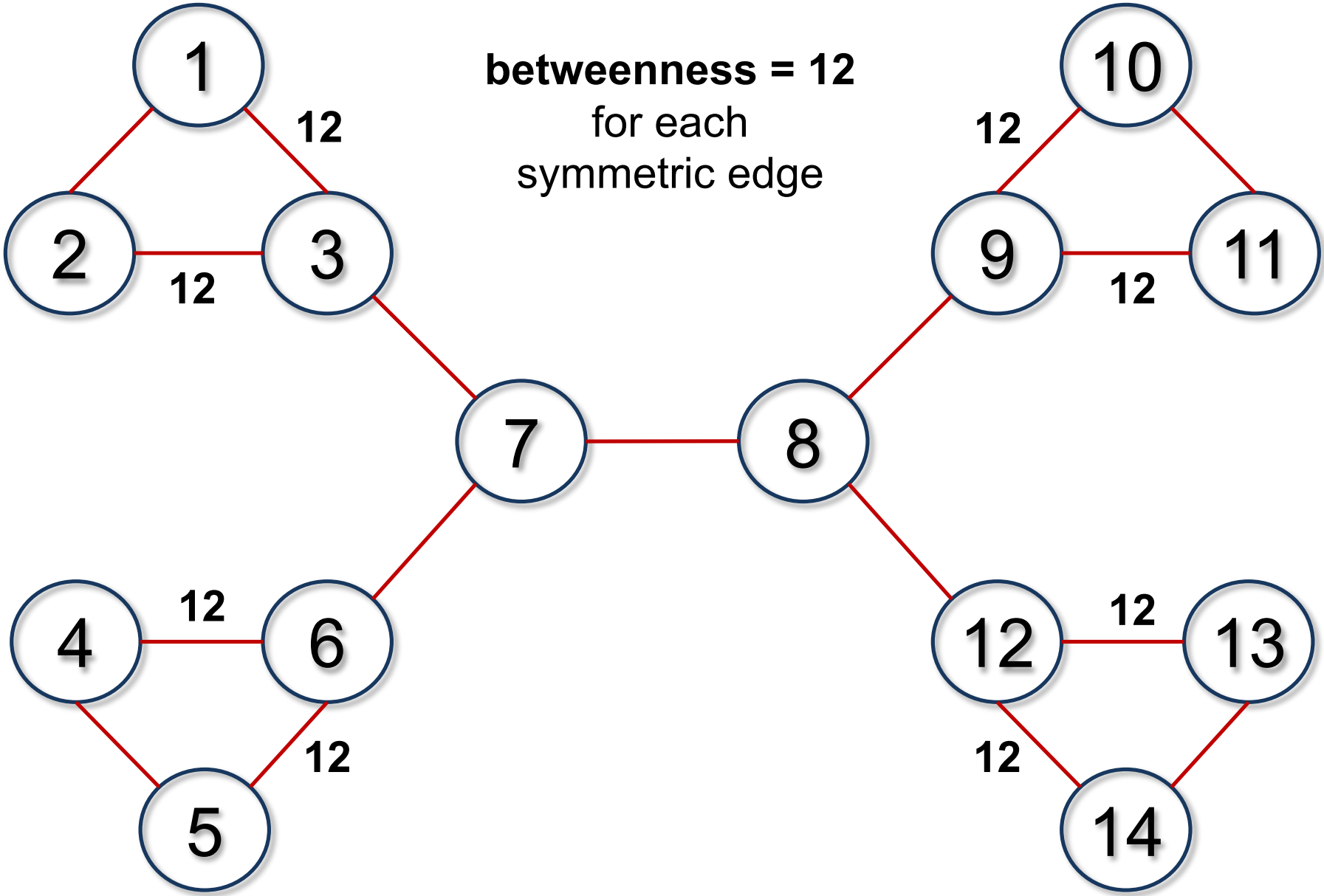

1
2
3
4
5
6
7
8
9
10
11
12
13
14
Calculate
betweenness for edge
3-7

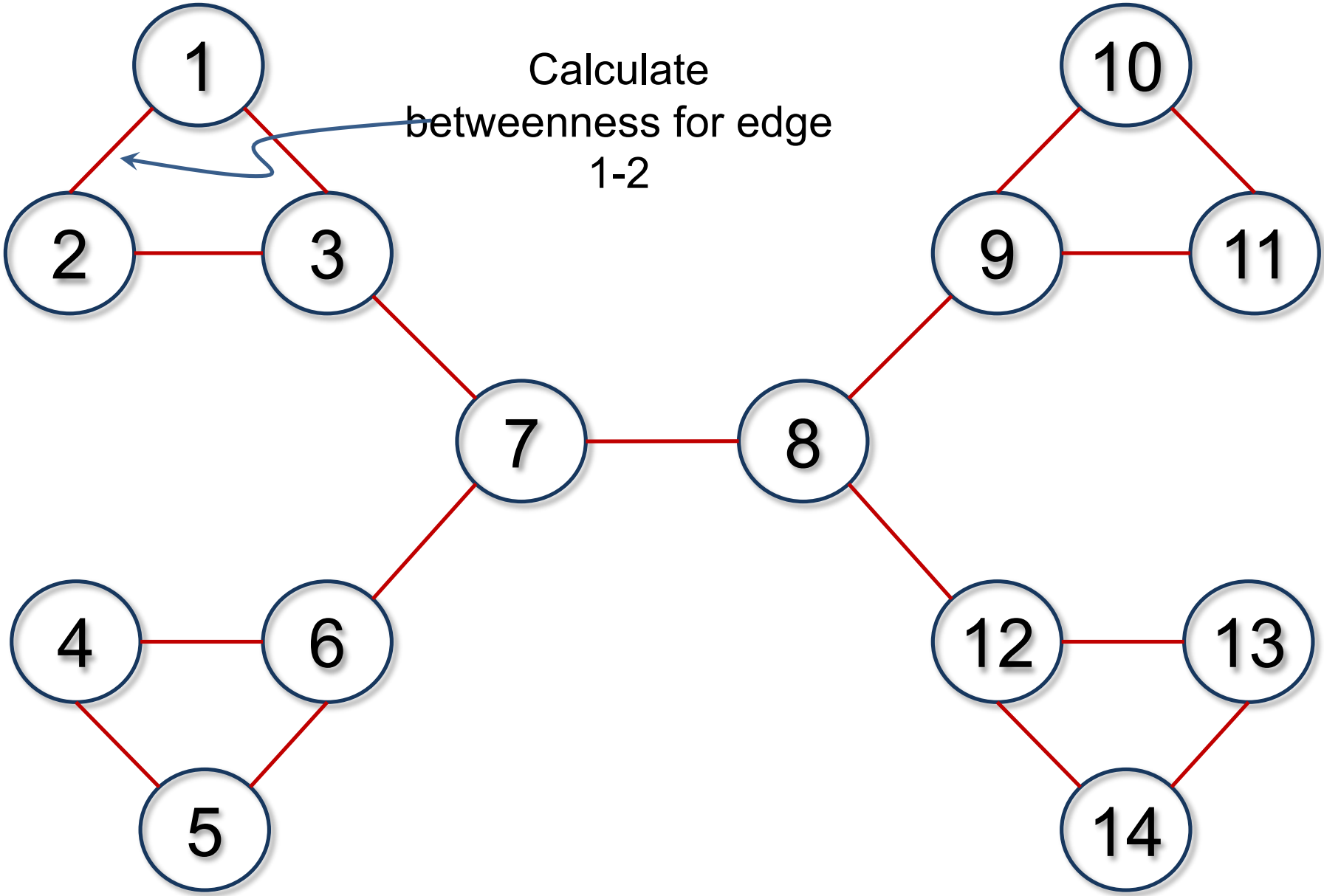

1
2
3
10
9
11
7
8
4
6
12
13
5
14
3 units flow from
1-3 to each 4-14
node,
so total =
3 x 11 = **33**

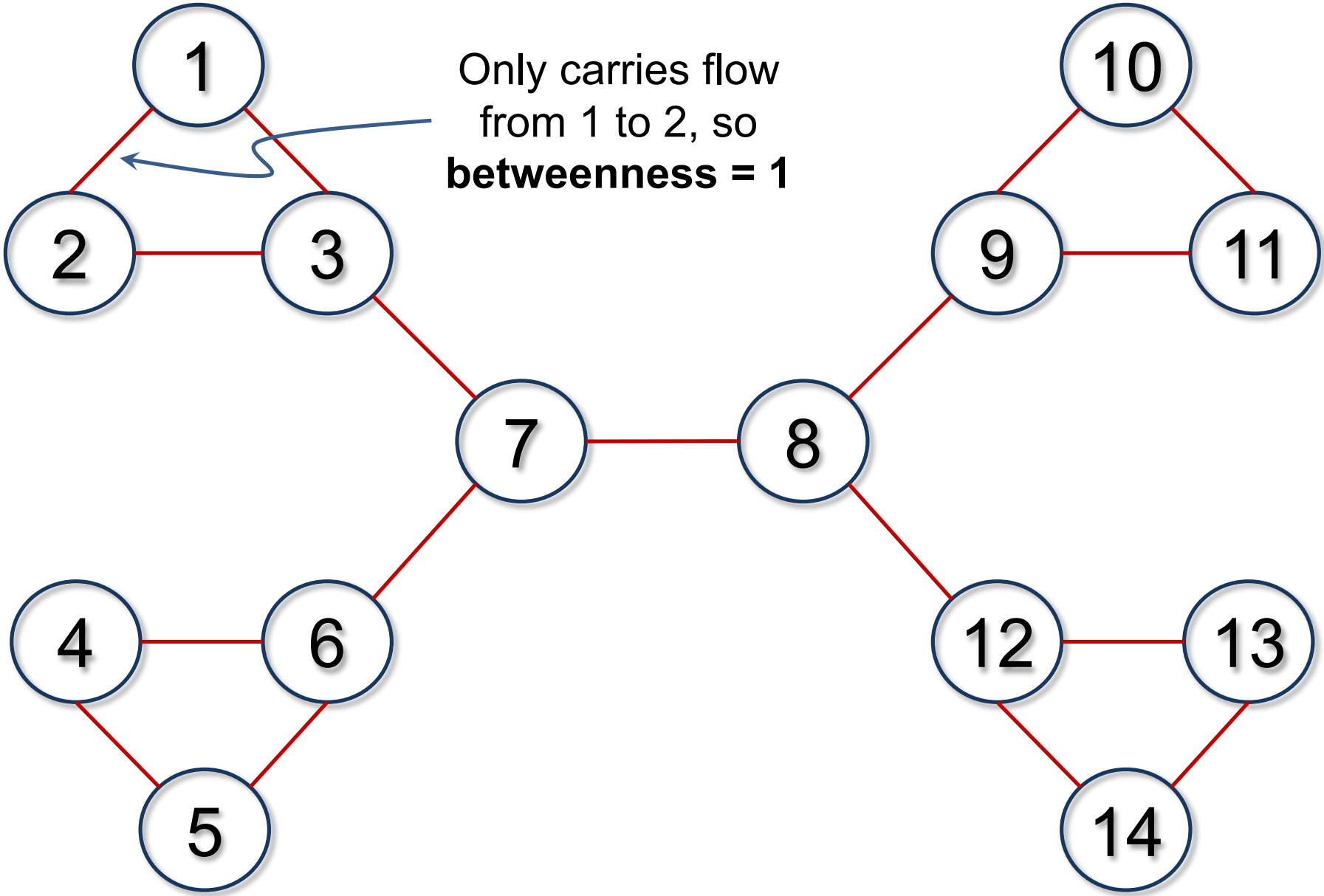

Betweenness = 33
for each
symmetric edge
1
2
3
4
5
6
7
8
9
10
11
12
13
14
33
33
33
33

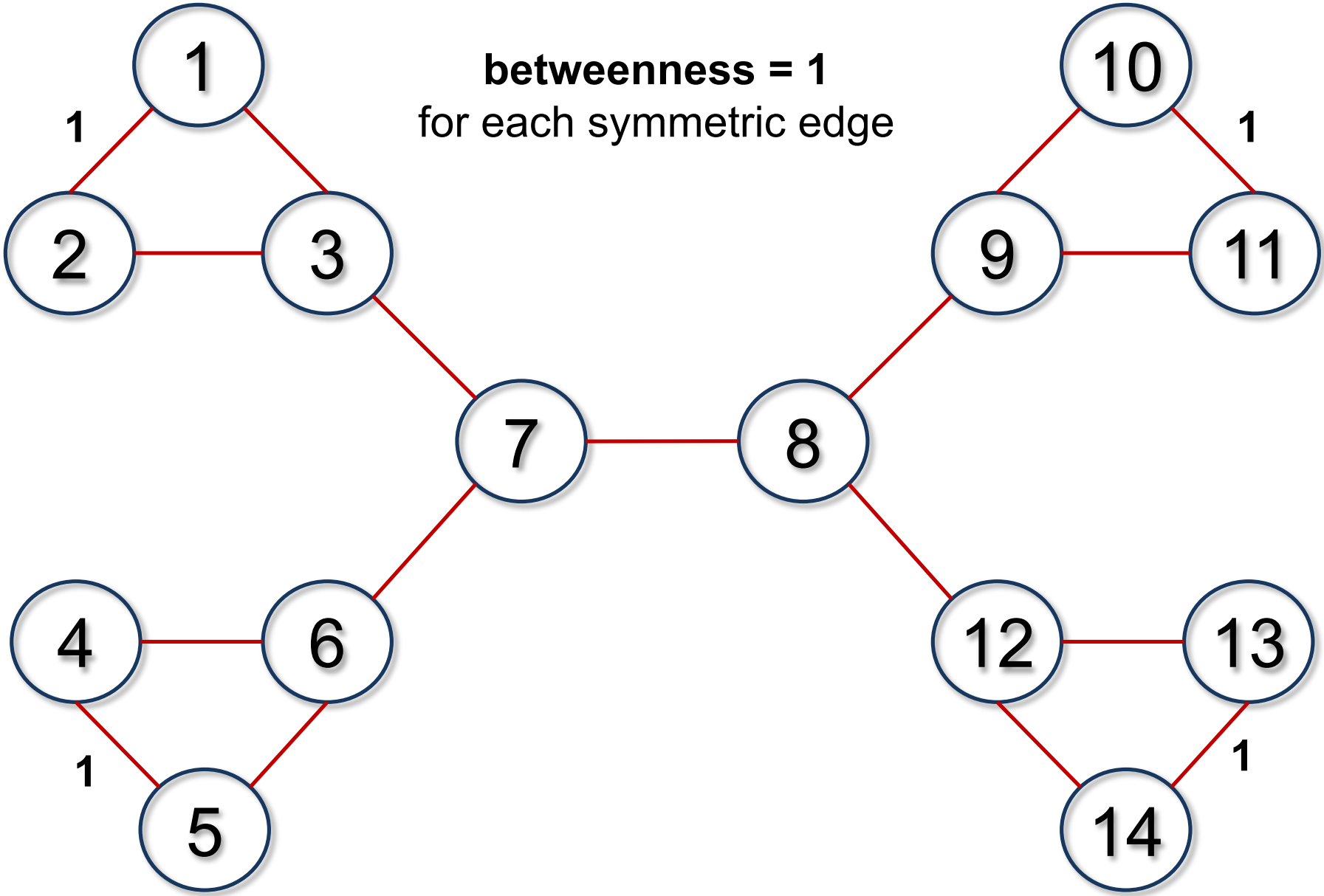

Calculate
betweenness for edge
1-3
1
2
3
4
5
6
7
8
9
10
11
12
13
14

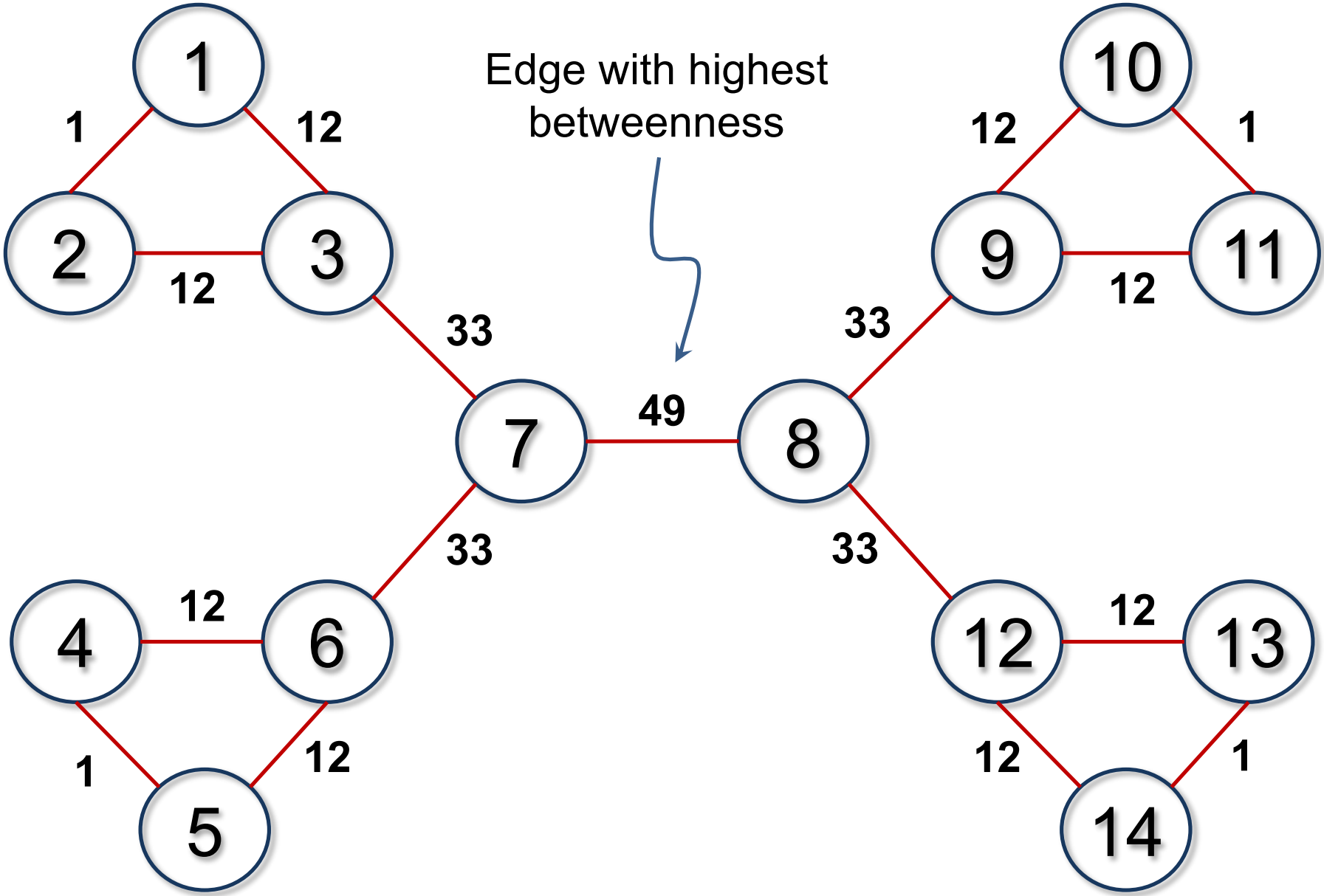

1
2
3
4
5
6
7
8
9
10
11
12
13
14
Carries all flow to node 1 except from node 2, so **betweenness = 12**


betweenness = 12
for each
symmetric edge
1
2
3
4
5
6
7
8
9
10
11
12
13
14
12
12
12
12
12
12
12
12

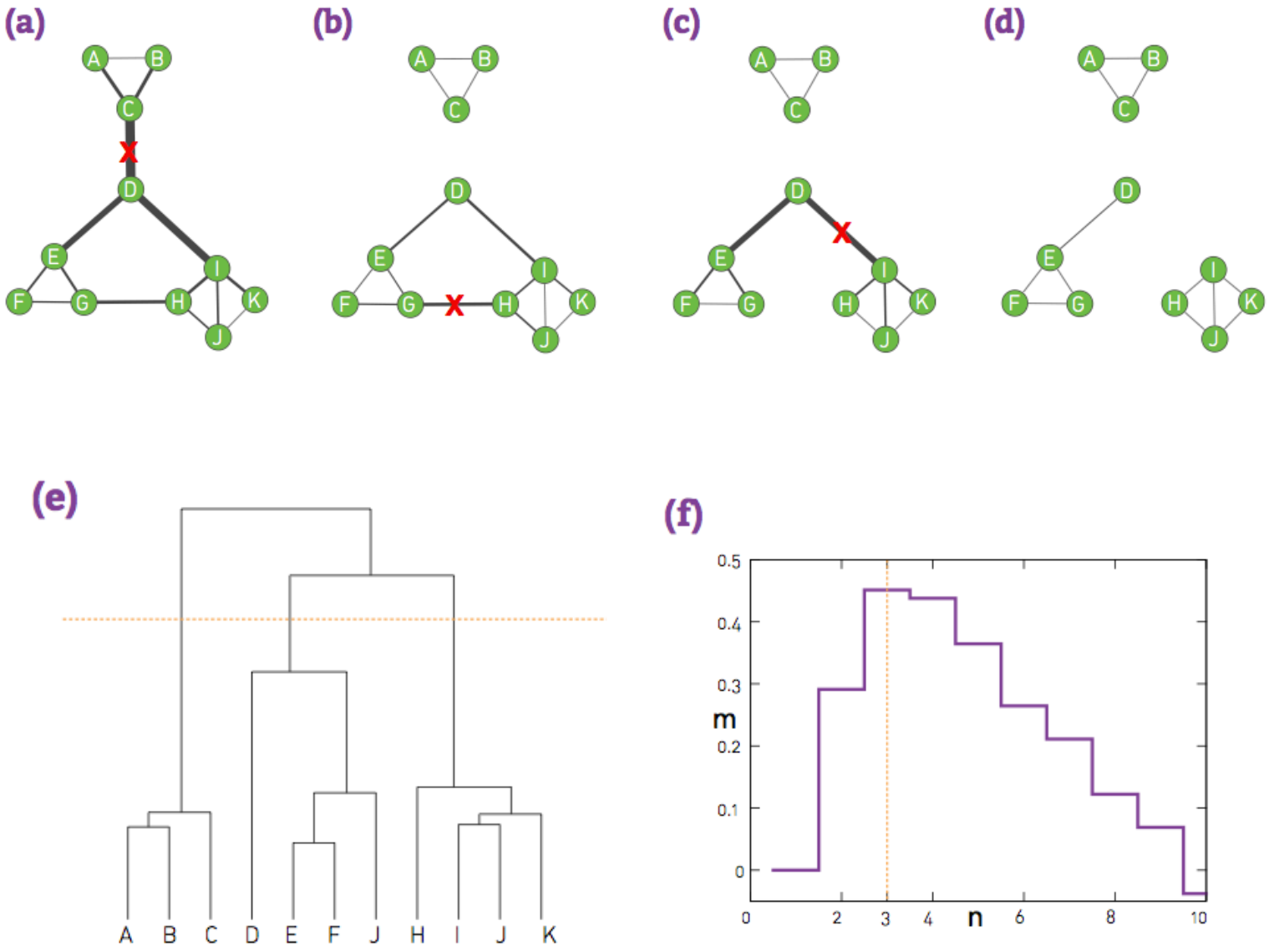

Calculate
betweenness for edge
1-2
1
2
3
4
5
6
7
8
9
10
11
12
13
14


1
2
3
4
5
6
7
8
9
10
11
12
13
14
Only carries flow
from 1 to 2, so
betweenness = 1


betweenness = 1
for each symmetric edge
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1
1
1
1


Edge with highest betweenness
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1
12
12
33
33
49
33
33
12
12
1
12
12
12
1

Hierarchical Clustering: compute centrality of each link; remove link with highest centrality; recalculate centrality; build dendrogram; choose communities that maximizes **modularity**;



quantifying quality of community structure | Modularity

How to select the number of clusters/evaluate the algorithm?

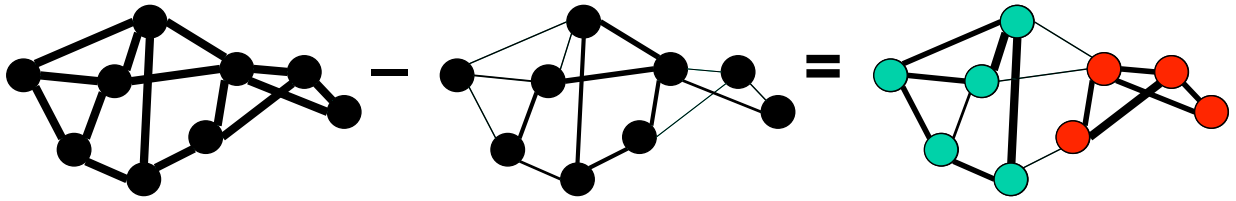
Random graphs are not expected to have community structure, so we will use them as null models.

$Q =$ (nr. of intra-cluster communities) $-$ (expected nr of edges)

In particular:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij})\ \delta(C_i, C_j)$$

where P_{ij} is the expected number of edges between nodes i and j under the null model, C_i is the community of vertex i, and $\delta(C_i, C_j) = 1$ if $C_i = C_j$ and 0 otherwise.



quantifying quality of community structure | Modularity

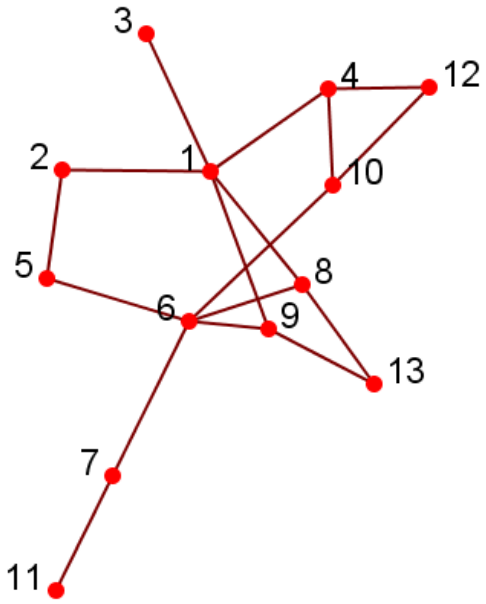
How to computer $\mathbf{P_{ij}}$?

The "configuration" random graph model choses a graph with the same degree distribution as the original graph uniformly at random.

- Let us compute P_{ij}
- There are $2m$ *stubs* or half-edges available in the configuration model
- Let p_i be the probability of picking at random a stub incident with i

$$p_i = \frac{k_i}{2m}$$

- The probability of connecting i to j is then $p_i p_j = \frac{k_i k_j}{4m^2}$
- And so $P_{ij} = 2m p_i p_j = \frac{k_i k_j}{2m}$

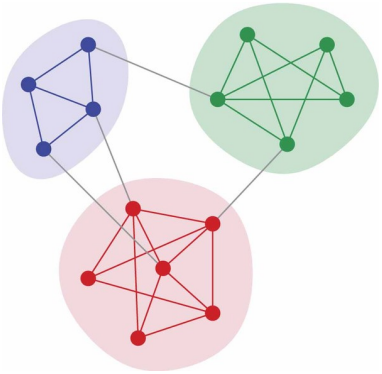


Expected Number of edges between 6 and 9 is
5*3/(2*17) = 15/34

quantifying quality of community structure | Modularity

Let n_c - number of classes, c_i - class label per node
Compare fraction of edges within the cluster to expected fraction if edges were distributed at random
Modularity:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j), \quad \delta(c_i, c_j)\text{- kronecker delta}$$



Q = (# edges within group s) – (expected # edges within group s)

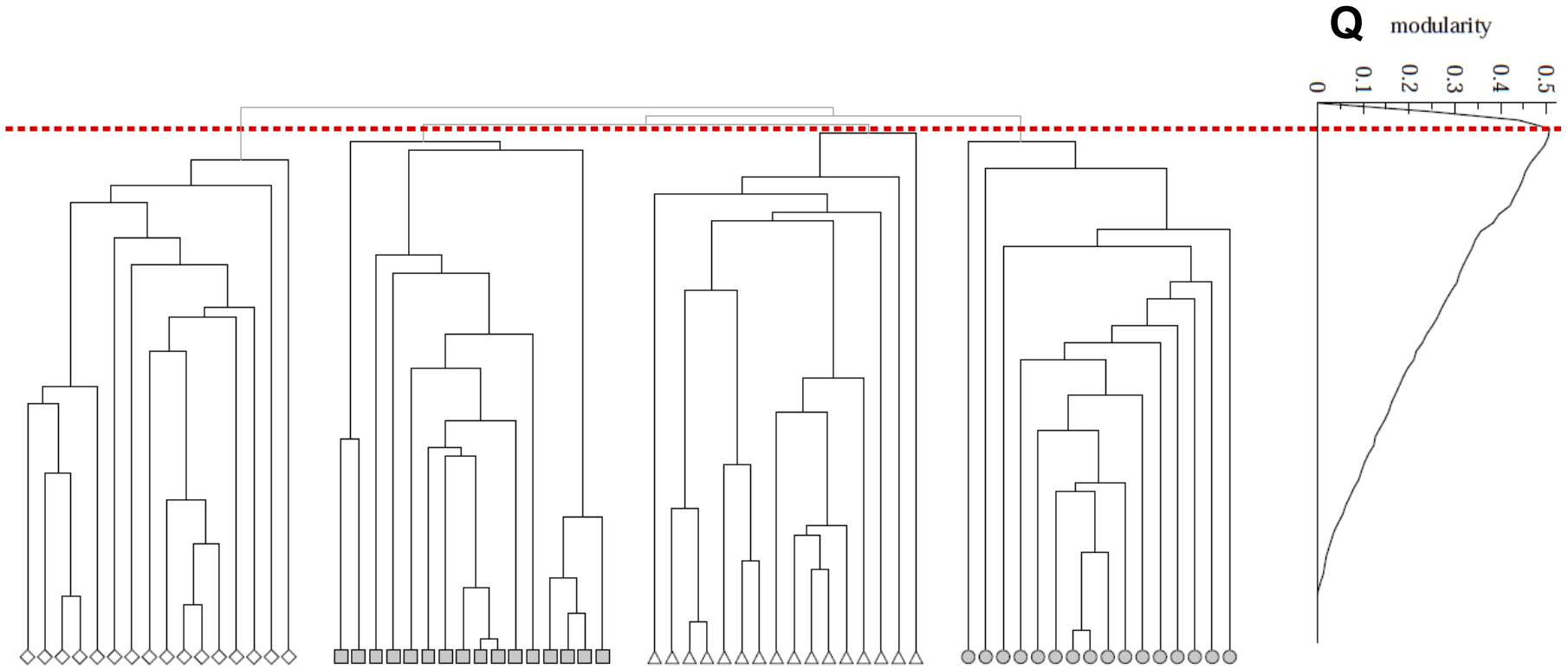
Positive Q means the number of edges within groups exceeds the expected number

The higher the modularity score - the better is community
Modularity score range $Q \in [-1/2, 1)$
Single class, $\delta(c_i, c_j) = 1$, $Q = 0$

quantifying quality of community structure | Modularity

Useful for selecting number of clusters;

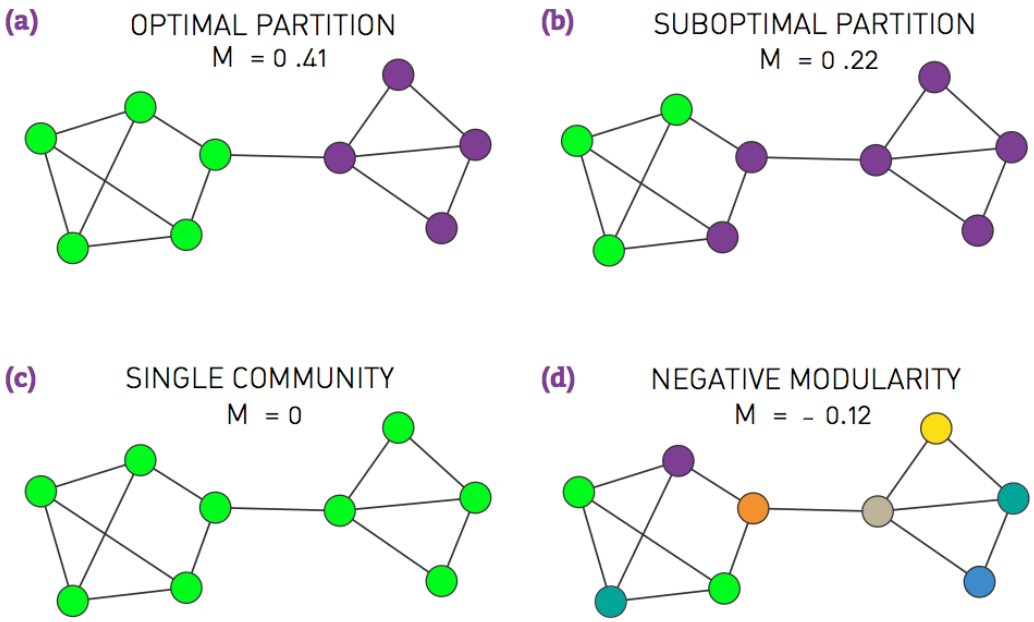
Modularity can be optimized directly (e.g. Louvain algorithm, Spectral algorithm);



quantifying quality of community structure | Modularity

Optimization

Which partition $\{C_c,\ c = 1,\ n_c\}$ **?**



- *Optimal partition*, that maximizes the modularity.
- *Sub-optimal* but positive modularity.
- *Negative Modularity*: If we assign each node to a different community.
- *Zero modularity:* Assigning all nodes to the same community, independent of the network structure.
- *Modularity is size dependent*

quantifying quality of community structure | Modularity

Optimization

A *greedy algorithm*, which iteratively joins nodes if the move increases the new partition's modularity.

Step 1. Assign each node to a community of its own. Hence we start with N communities.

Step 2. Inspect each pair of communities connected by at least one link and compute the modularity variation obtained if we merge these two communities.

Step 3. Identify the community pairs for which ΔM is the largest and merge them. Note that modularity of a particular partition is always calculated from the full topology of the network.

Step 4. Repeat step 2 until all nodes are merged into a single community.

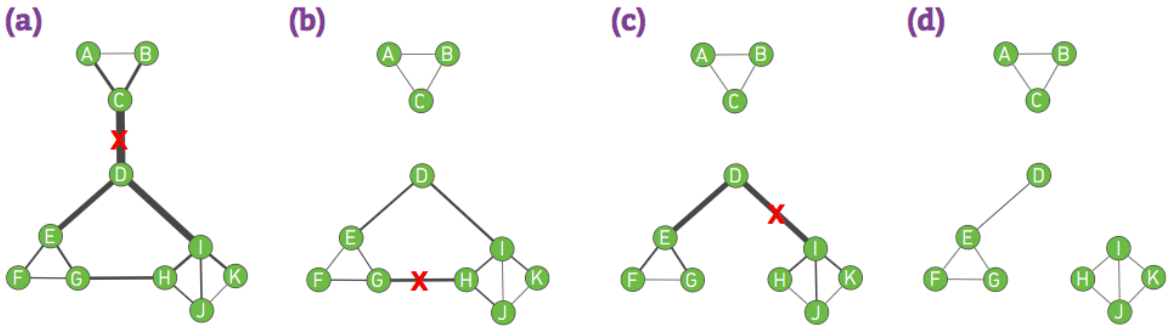
Step 5. Record for each step and select the partition for which the modularity is maximal.

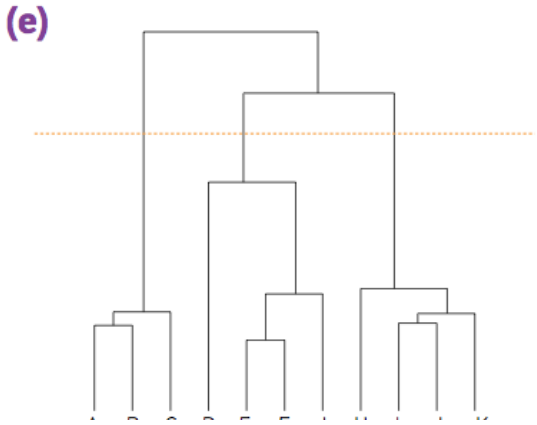
quantifying quality of community structure | Modularity

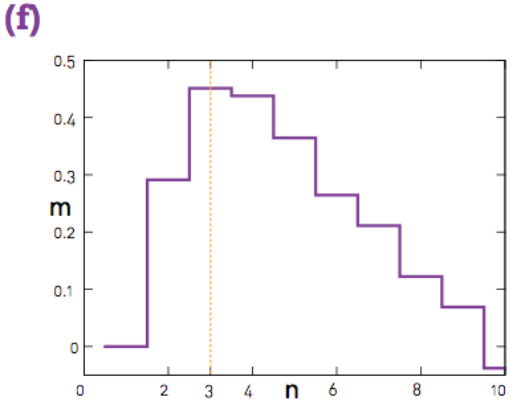
Optimization

Which partition $\{C_c,\ c=1,\ n_c\}$ **?**

$$M(C_c) = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$







Random Networks

Erdös-Rényi Random Network

Definition:

A **random graph** is a graph of N nodes where each pair of nodes is connected by probability **p**.

$G(N, L)$ Model

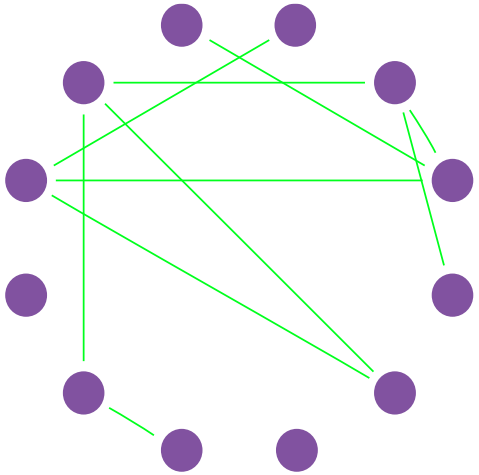
N labeled nodes are connected with L randomly placed links. Erdős and Rényi used this definition in their string of papers on random networks [2-9].

$G(N, p)$ Model

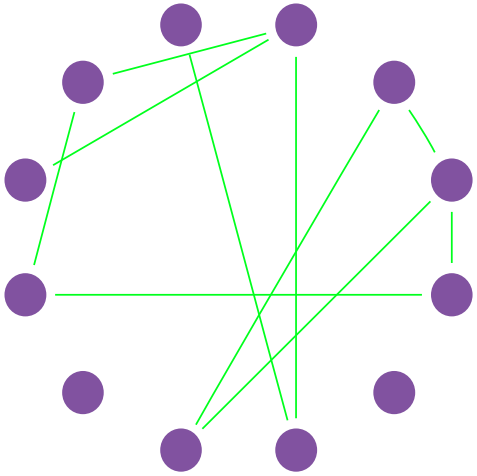
Each pair of N labeled nodes is connected with probability p, a model introduced by Gilbert [10].

The number of links is variable

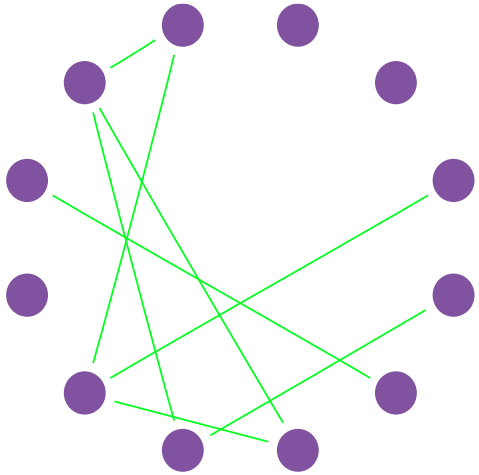
p=1/6
N=12



L=8

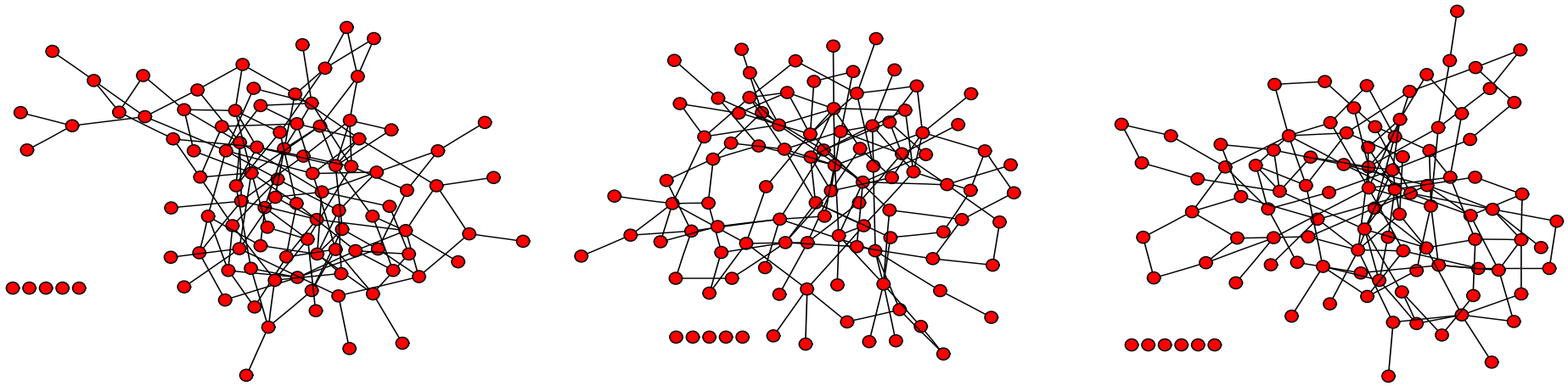


L=10



L=7

p=0.03
N=100



Number of links in a random network

P(L): the probability to have exactly *L* links in a network of *N* nodes and probability *p*:

The maximum number of links in a network of N nodes.

$$P(L) = \binom{\binom{N}{2}}{L} p^L (1-p)^{\frac{N(N-1)}{2}-L}$$

Number of different ways we can choose L links among all potential links.

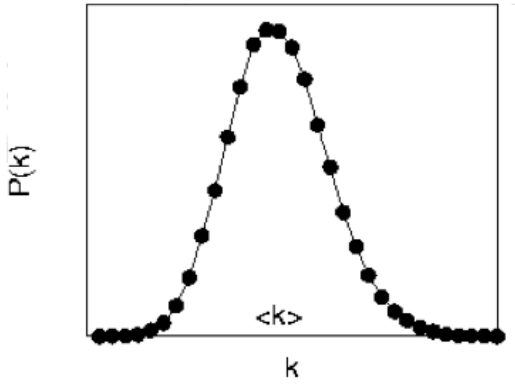
Binomial distribution...

•The average number of links <*L*> in a random graph

$$< L >= \sum_{L=0}^{\frac{N(N-1)}{2}} LP(L) = p\frac{N(N-1)}{2}$$

$$< k >= 2L/N = p(N-1)$$

DEGREE DISTRIBUTION OF A RANDOM GRAPH



$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k}$$

Select k nodes from N-1

probability of having *k* edges

probability of missing N-1-k edges

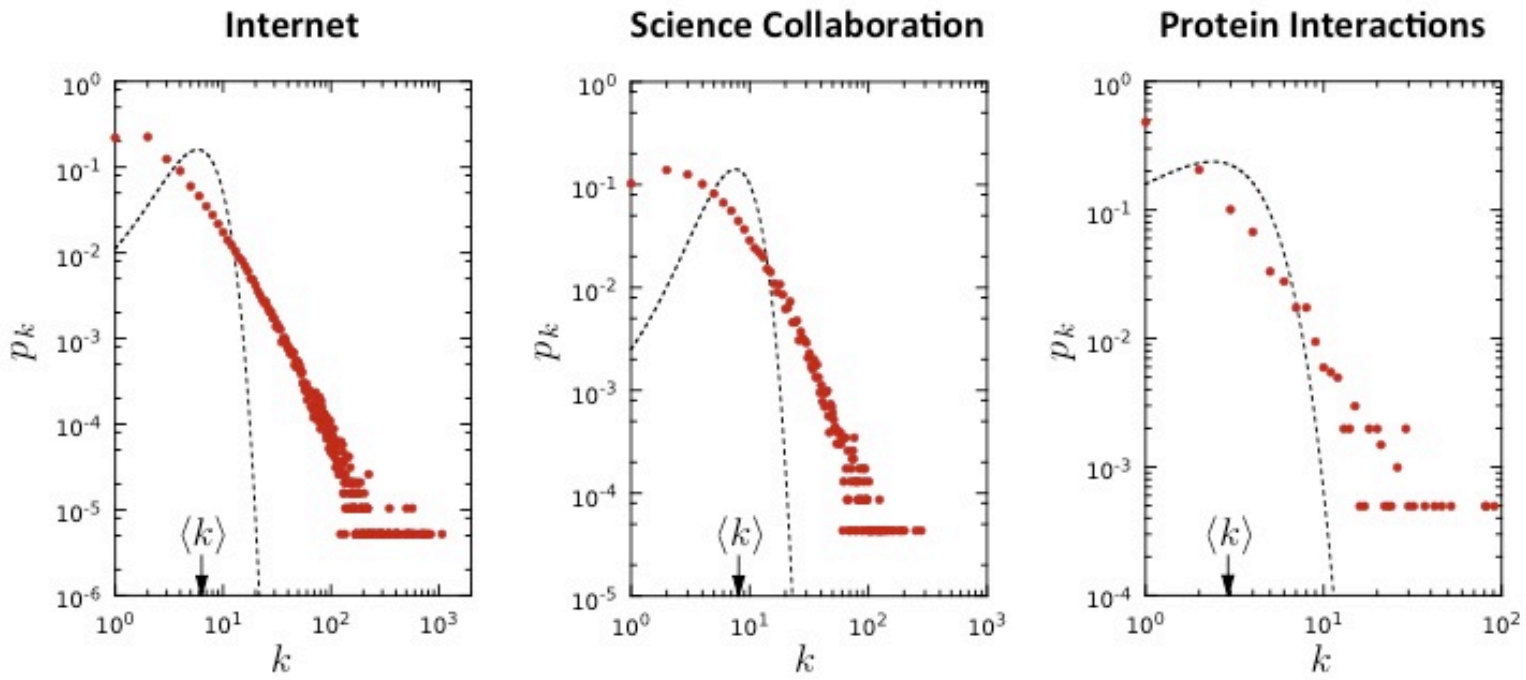
$$< k >= p(N-1) \qquad \sigma_k^2 = p(1-p)(N-1)$$

$$\frac{\sigma_k}{< k >} = \left[\frac{1-p}{p}\frac{1}{(N-1)}\right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

As the network size increases, the distribution becomes increasingly narrow—we are increasingly confident that the degree of a node is in the vicinity of <k>.

Insights: we don't expect large hubs in the network

Real Networks are not Poisson

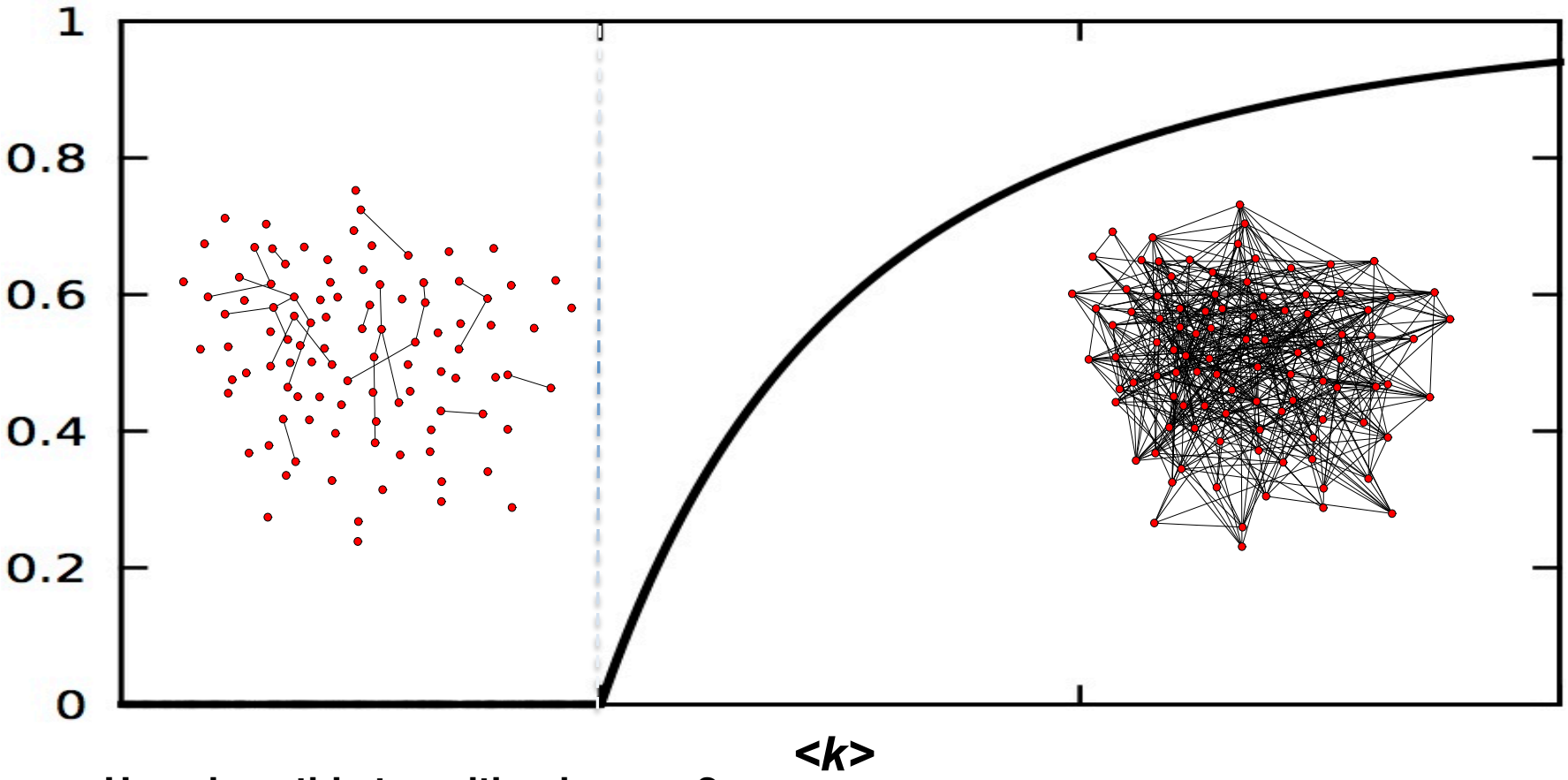


Phase transition of the size of the giant component in the Erdös-Rényi Random Network

- The largest component in the ER random graph has constant size 1 when $p = 0$ and extensive size n when $p = 1$.
- An interesting question to ask is how the transition between these two extremes occurs if we construct random graphs with gradually increasing values of p, starting at 0 and ending up at 1—this is bond percolation!
- It turns out that the size of the largest component undergoes a sudden change, or phase transition, from constant size to extensive size at one particular special value of $p = 1/n$.

EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK**.



How does this transition happen?

EVOLUTION OF A RANDOM NETWORK

disconnected nodes ➔ **NETWORK**.

$\langle k_c \rangle = 1$ (Erdos and Renyi, 1959)

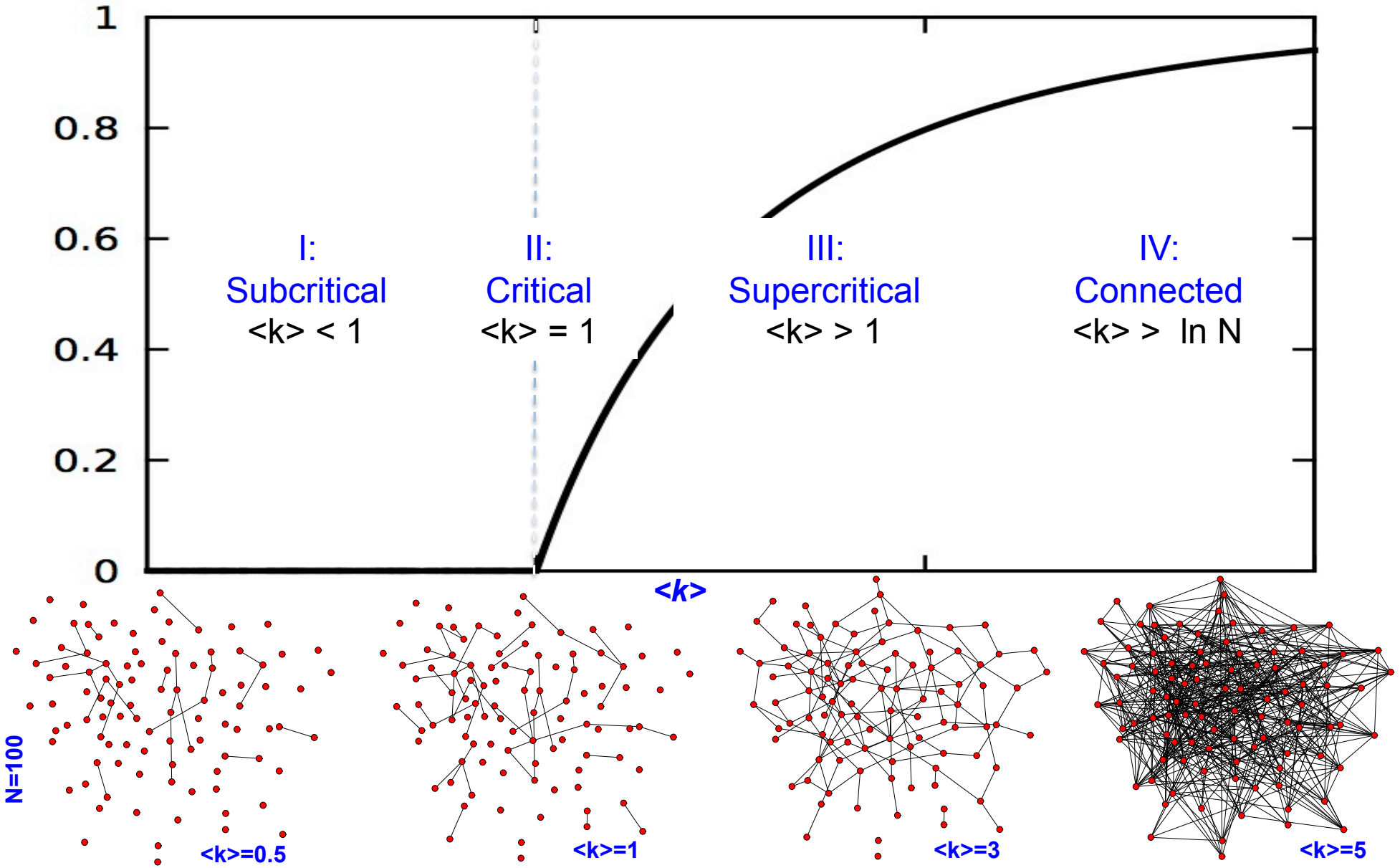
The fact that at least one link per node is *necessary* to have a giant component is not unexpected. Indeed, for a giant component to exist, each of its nodes must be linked to at least one other node.

It is somewhat unexpected, however that one link is *sufficient* for the emergence of a giant component.

It is equally interesting that the emergence of the giant cluster is not gradual, but follows what physicists call a second order phase transition at <k>=1.

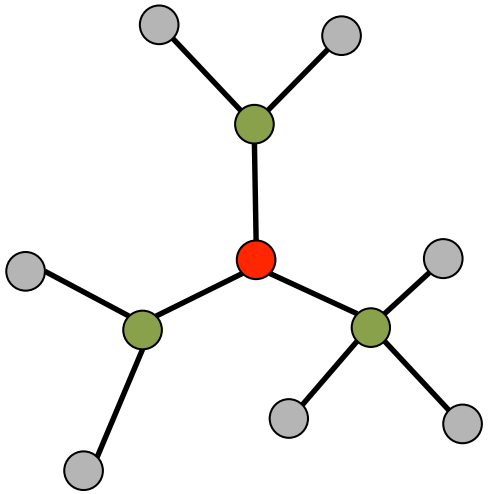
The size of the giant component in the Erdös-Rényi Random Network (Bollobás et al., 2001)

- If $p < \frac{1}{n}$
  - with high probability, there is no giant component, with all connected components of the graph having size $O(\log n)$.
- If $p > \frac{1}{n}$
  - with high probability, there is a single giant component, with all other components having size $O(\log n)$.
- If $p = \frac{1}{n}$
  - with high probability, the number of vertices in the largest component of the graph is proportional to $n^{2/3}$.


1
0.8
0.6
0.4
0.2
0
I:
Subcritical
<k> < 1
II:
Critical
<k> = 1
III:
Supercritical
<k> > 1
IV:
Connected
<k> > ln N
<k>
N=100
<k>=0.5
<k>=1
<k>=3
<k>=5

DISTANCES IN RANDOM GRAPHS

Random graphs tend to have a tree-like topology with almost constant node degrees.



<k> nodes at distance one (d=1).

<k>^{2} nodes at distance two (d=2).

<k>^{3} nodes at distance three (d =3).

...

<k>^{d} nodes at distance d.

$$N = 1 + \langle k \rangle + \langle k \rangle^2 + \ldots + \langle k \rangle^{d_{\max}} = \frac{\langle k \rangle^{d_{\max}+1} - 1}{\langle k \rangle - 1} \approx \langle k \rangle^{d_{\max}} \quad \Rightarrow \quad d_{\max} = \frac{\log N}{\log \langle k \rangle}$$

DISTANCES IN RANDOM GRAPHS

$$d_{\max} = \frac{\log N}{\log \langle k \rangle}$$

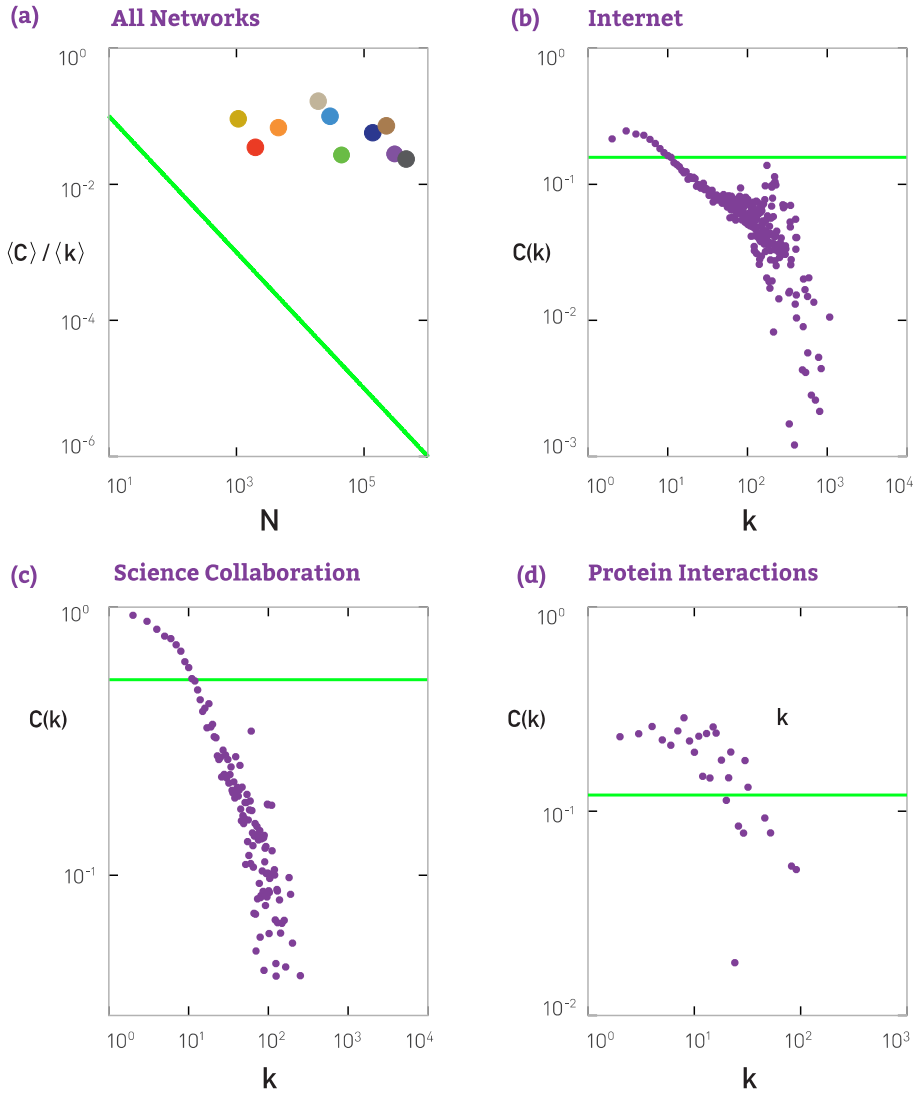
In most networks this offers a better approximation to the average distance between two randomly chosen nodes, ⟨d⟩, than to d_{max} .

$$< d >= \frac{\log N}{\log \langle k \rangle}$$

We will call the *small world phenomena* the property that the average path length or the diameter depends logarithmically on the system size. Hence, ”small” means that ⟨d⟩ is proportional to log N, rather than N.

The 1/log⟨k⟩ term implies that denser the network, the smaller will be the distance between the nodes.

CLUSTERING COEFFICIENT



$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$

C decreases with the system size *N*.

C is independent of a node's degree *k*.

- Clustering coefficient distribution is hard to find. So we focus on the expectation.
- The average Clustering coefficient in the random network is approximately

$$\langle C \rangle \approx \frac{\langle K \rangle}{n}$$

- Randomly select a node i, there are k_i friends, leading to $k_i(k_i - 1)/2$ maximum possible edges, and each will appear with probability p. So the average

$$\langle C \rangle = p \approx \frac{\langle K \rangle}{n}$$

Characteristics of a **random network**

- Sparsity: Average density $= p$.
- Degree distribution: Poisson distribution

$$\begin{aligned} P(K = k) &= \binom{n}{k-1} p^k (1-p)^{n-k} \\ &\approx e^{-\langle K \rangle} \frac{\langle K \rangle^k}{k!}. \end{aligned}$$

- Average path: small world

$$\langle D \rangle \approx \frac{\log n}{\log \langle K \rangle}$$

- Average clustering coefficient: low for large network

$$\langle C \rangle = p \approx \frac{\langle K \rangle}{n}$$

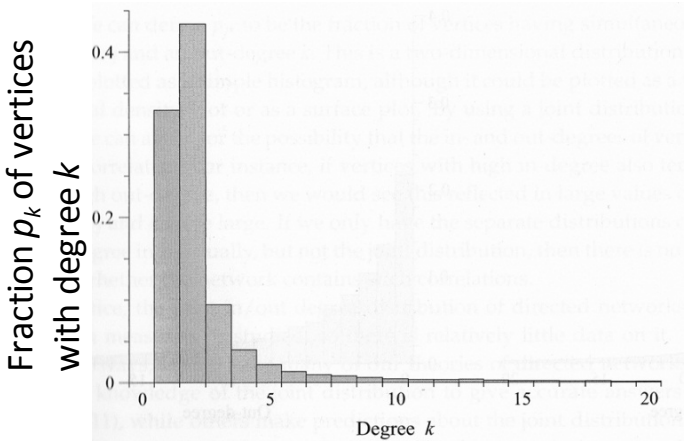
- The threshold for the emergence of the giant component is

$$p = \frac{1}{n} \text{ or } \langle K \rangle \approx 1$$

- No community structure
- No assortative mixing

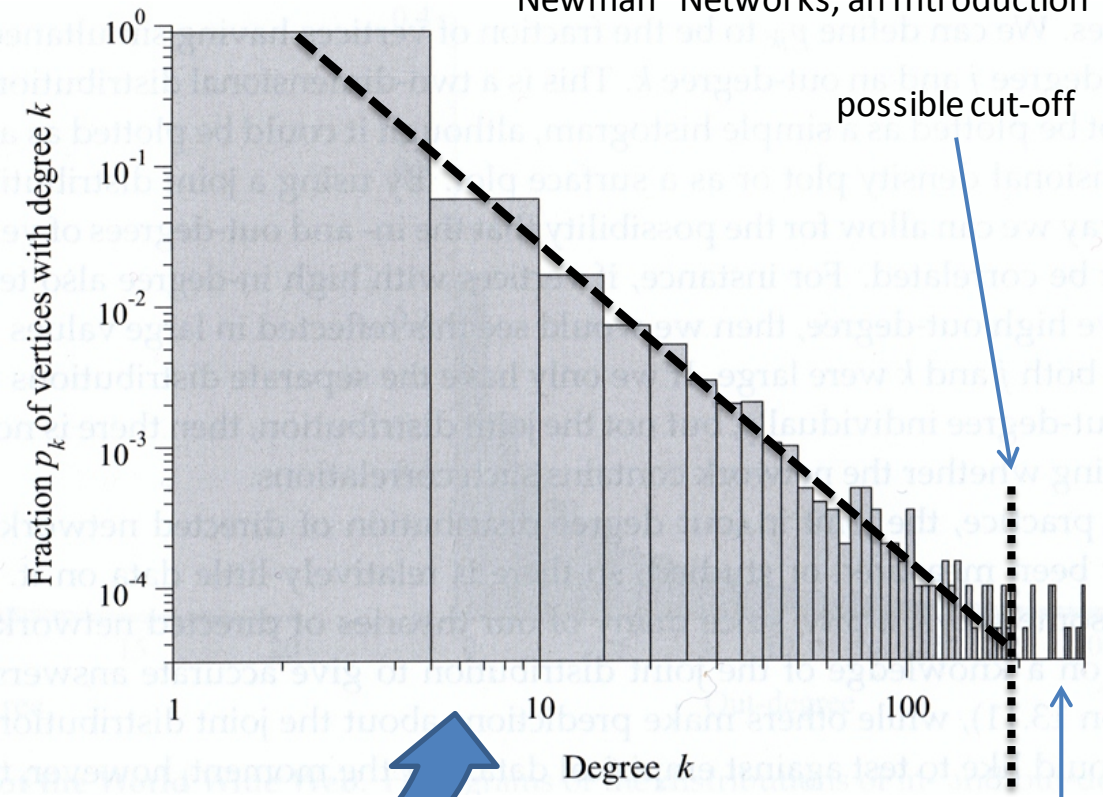
Real networks are not random

Power Laws (aka scale-free)



Internet at the level of autonomous systems

Newman "Networks, an Introduction"



logarithmic scales; bigger range of bins

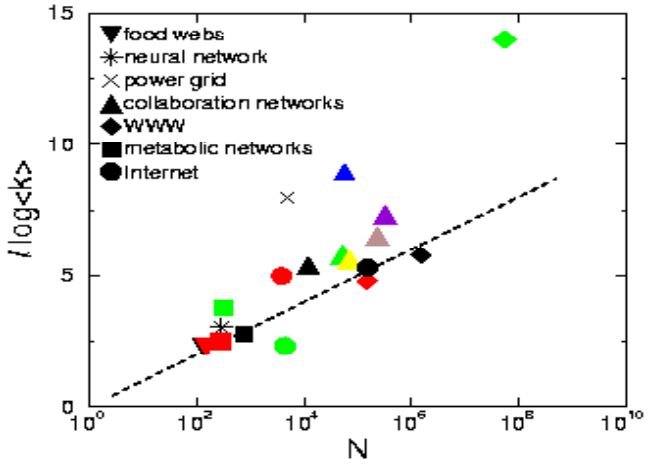
$$\ln p_k = -\alpha \ln k + c \text{ or } p_k = Ck^{-\alpha}, \text{ where } C = e^c$$
$$\text{typical } \alpha \in [2, 3] \text{ (see handout Table 8.1)}$$

Problem of histograms: statistics is poor at the tail of the distribution
Solution I: different sizes of bins

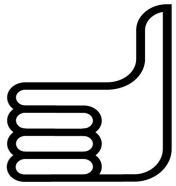
PATH LENGTHS IN REAL NETWORKS

Prediction:

$$< d >= \frac{\log N}{\log \langle k \rangle}$$



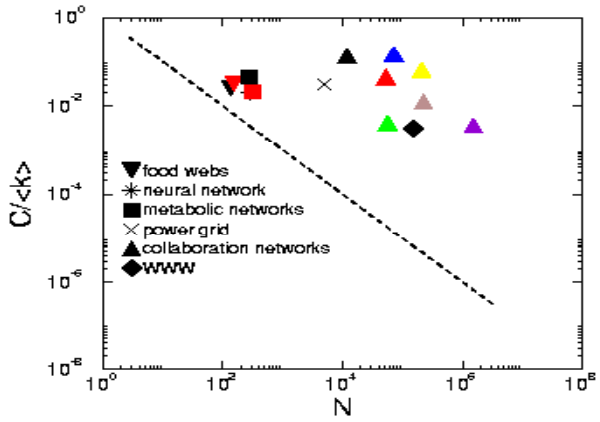
Real networks have short distances like random graphs.



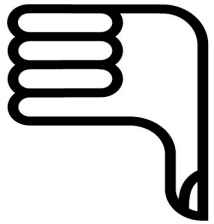
CLUSTERING COEFFICIENT

Prediction:

$$C_i = \frac{2\langle L_i \rangle}{k_i(k_i - 1)} = p = \frac{\langle k \rangle}{N}.$$



C_{rand} underestimates with orders of magnitudes the clustering coefficient of real networks.



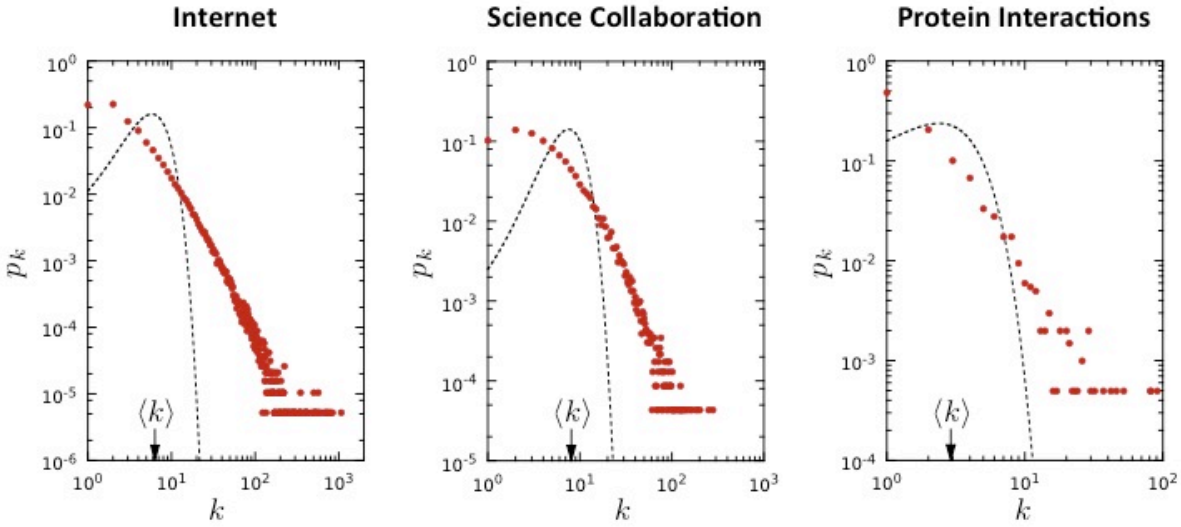
THE DEGREE DISTRIBUTION

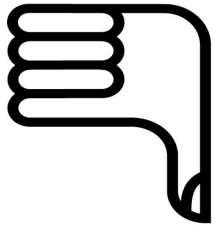
Prediction:

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$

Data:

$$P(k) \approx k^{-\gamma}$$





Characteristics of a **REAL network**

- Sparsity: $|E| = O(n)$ edges.
- Degree distribution: Power distribution (scale-free)
- Average path: $O(\log n)$, small world
- Average clustering coefficient: high for large network (compared to random network)
- Giant component: common
- Community structures: common
- Assortative mixing: common

ER network vs real network

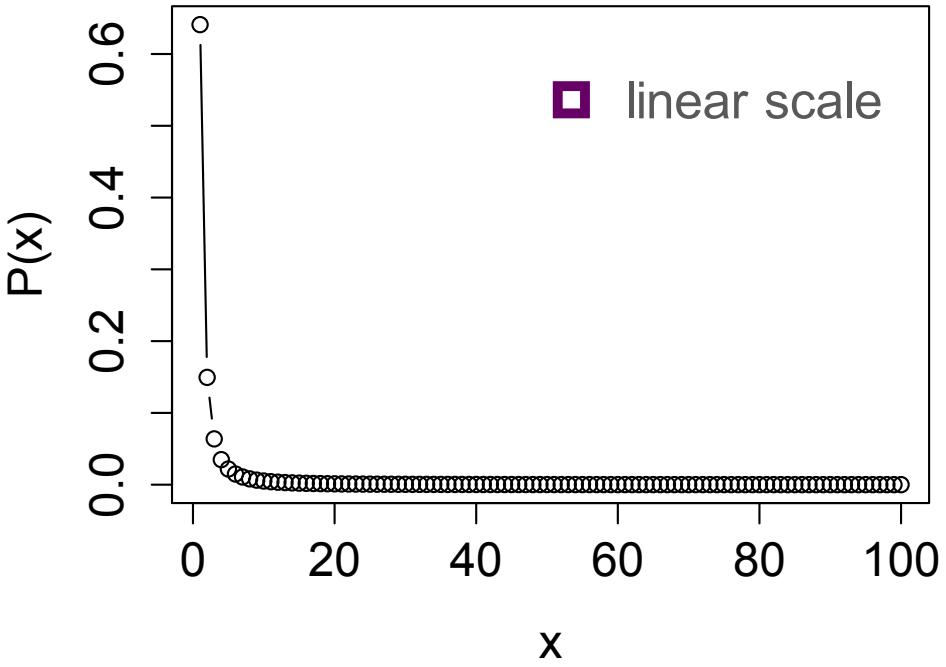
Characteristics	ER prediction	Real network
Density	$p \Longrightarrow$ Sparse	Sparse
Degree distribution	Poisson (or Normal)	Power-law
Clustering coefficient	$p \Longrightarrow$ Low	High
Average distance	Small world	Small world
Giant component	Yes	Yes
Community structure	No	Yes
Homophily	No	Yes

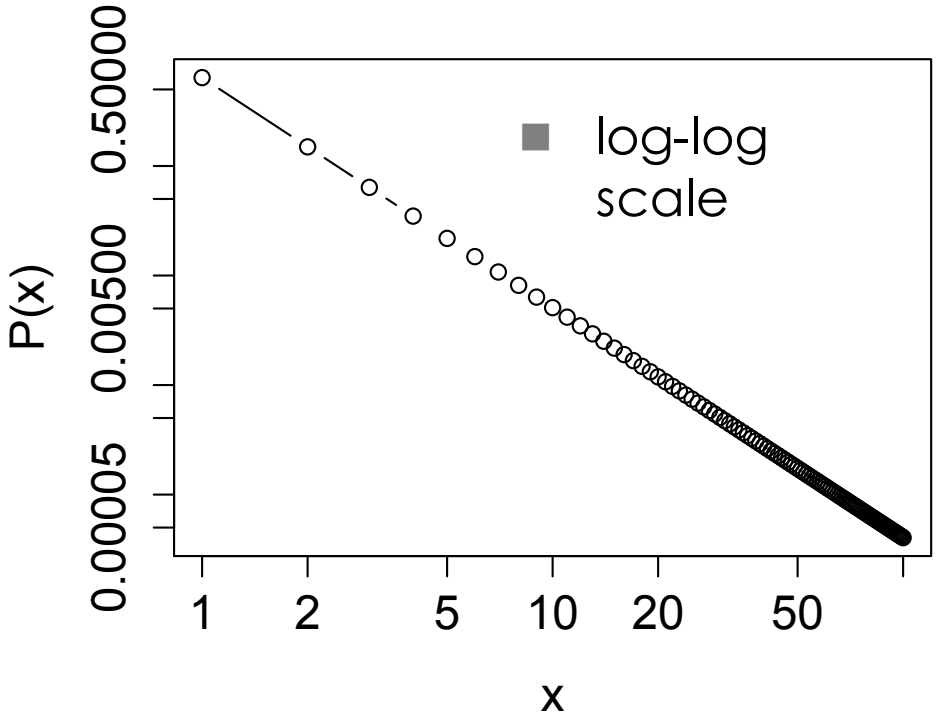
Two questions:

1. How to obtain **power-law distributions** from random network models?

2. How to obtain **higher cluster coefficients** from random network models?

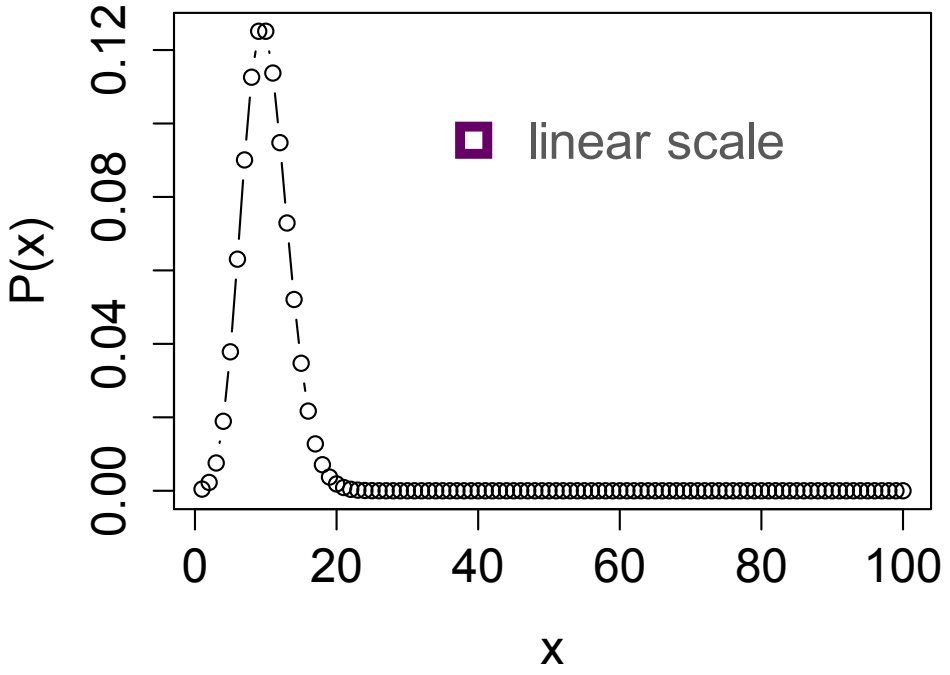
Power-law distribution

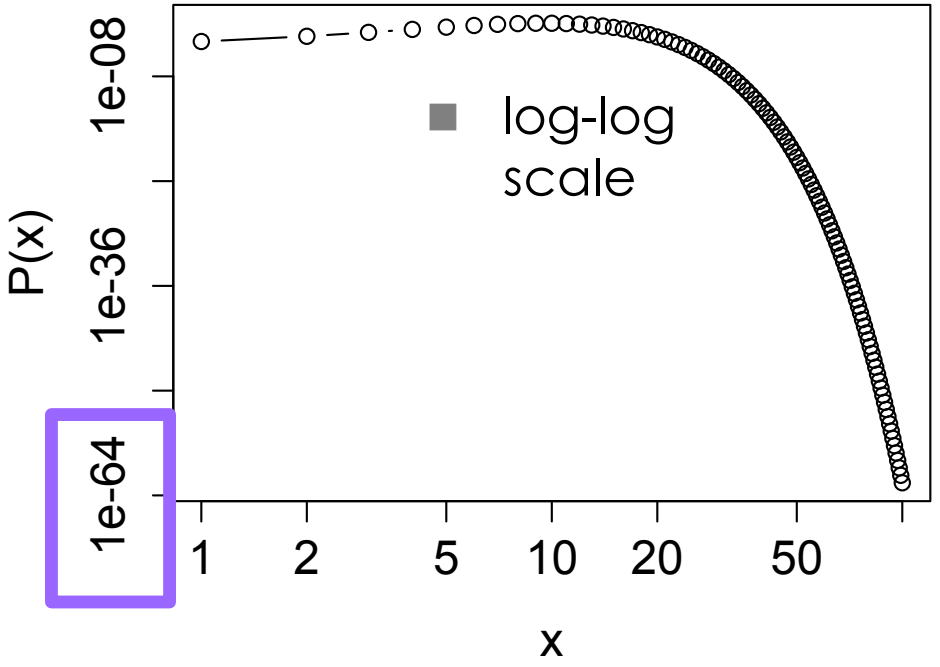




- high skew (asymmetry)
- straight line on a log-log plot

Poisson distribution

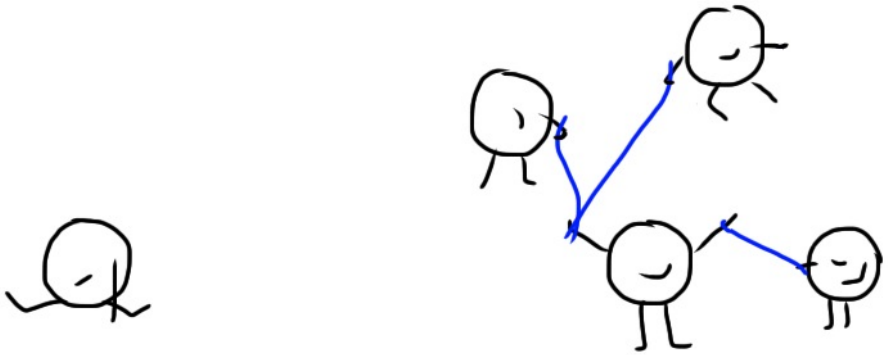




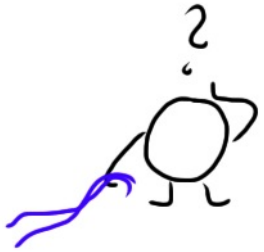
- little skew (asymmetry)
- curved on a log-log plot

2 ingredients in generating power-law networks

- nodes prefer to attach to nodes with many connections (preferential attachment, cumulative advantage)



- Process also known as
  - cumulative advantage
  - rich-get-richer
  - Matthew effect



Barabasi-Albert model

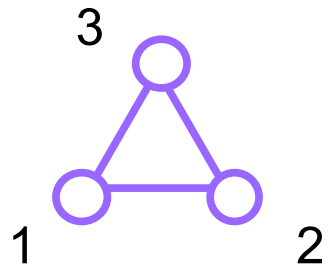
- First used to describe skewed degree distribution of the World Wide Web
- Each node connects to other nodes with probability proportional to their degree
  - the process starts with some initial subgraph
  - each new node comes in with *m* edges
  - probability of connecting to node *i*

$$\Pi(i) = m \frac{k_i}{\sum_j k_j}$$

- Results in power-law with exponent α = 3

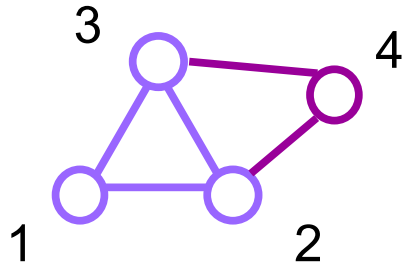
- To start, each vertex has an equal number of edges (2)
  - the probability of choosing any vertex is 1/3

1 1 2 2 3 3

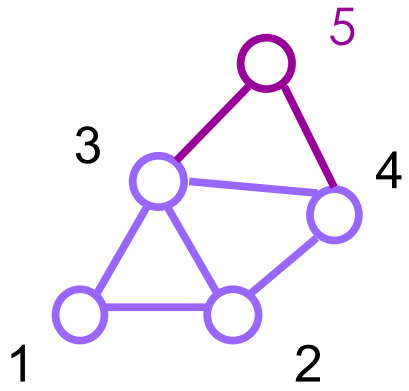


- We add a new vertex, and it will have m edges, here take m=2
  - draw 2 random elements from the array – suppose they are 2 and 3

1 1 2 2 2 3 3 3 4 4



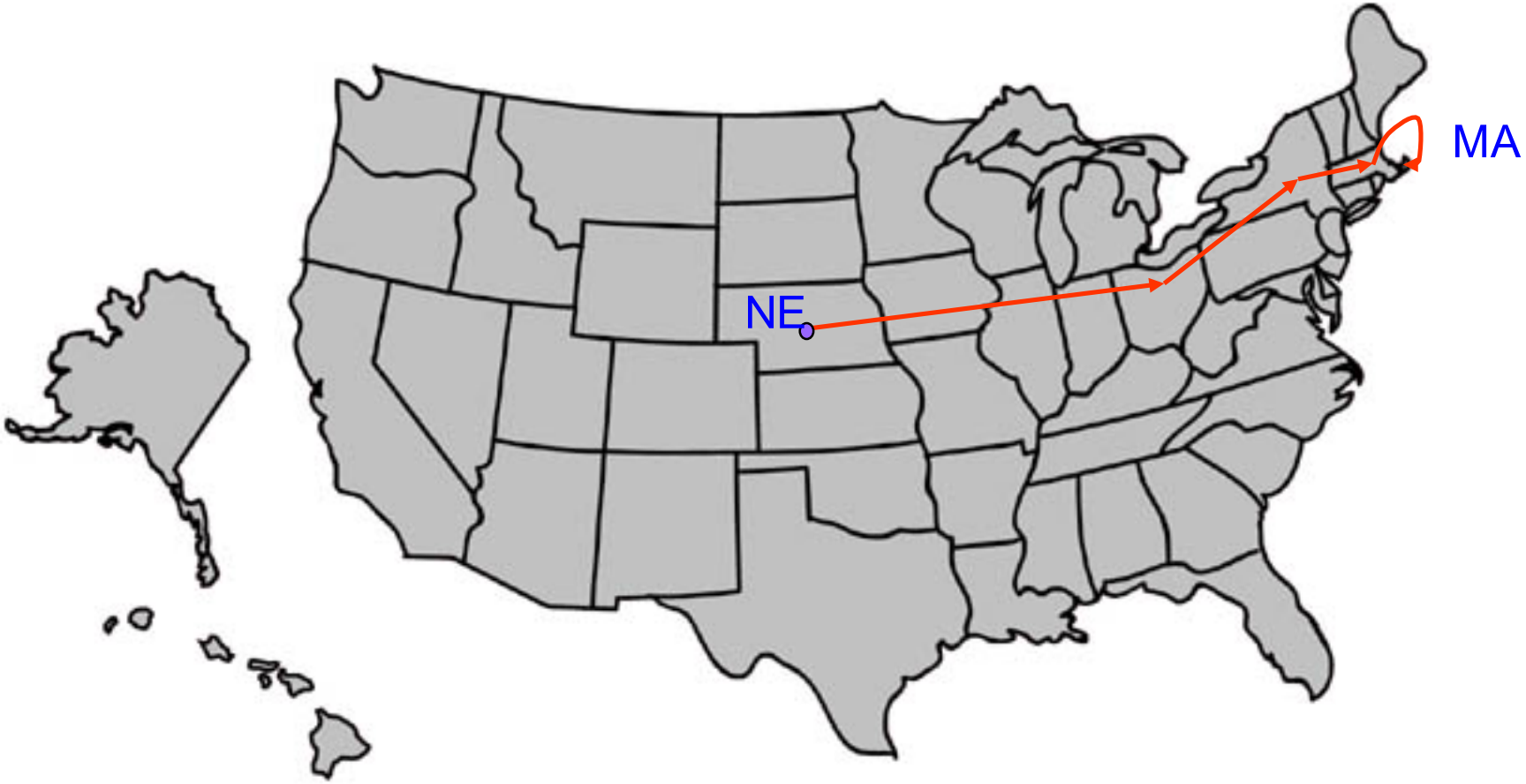
- Now the probabilities of selecting 1,2,3,or 4 are
  1/5, 3/10, 3/10, 1/5



- Add a new vertex, draw a vertex for it to connect from the array
  - etc.

1 1 2 2 2 3 3 3 3 4 4 4 5 5

Small world phenomenon:
Milgram's experiment



Milgram's experiment

Instructions:
Given a target individual (stockbroker in Boston), pass the message to a person you correspond with who is "closest" to the target.

Outcome:

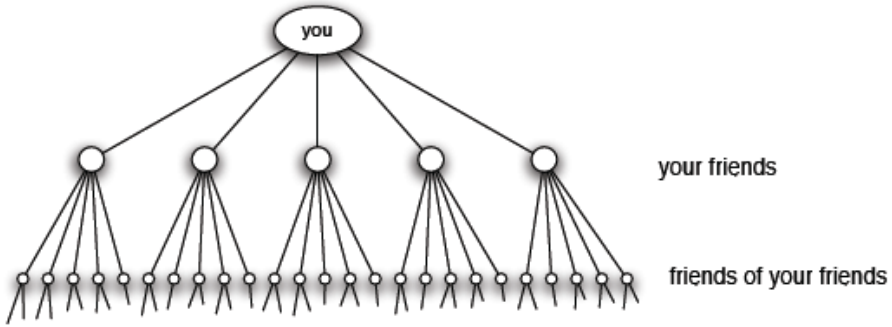
20% of initiated chains reached target
average chain length = 6.5

- "Six degrees of separation"

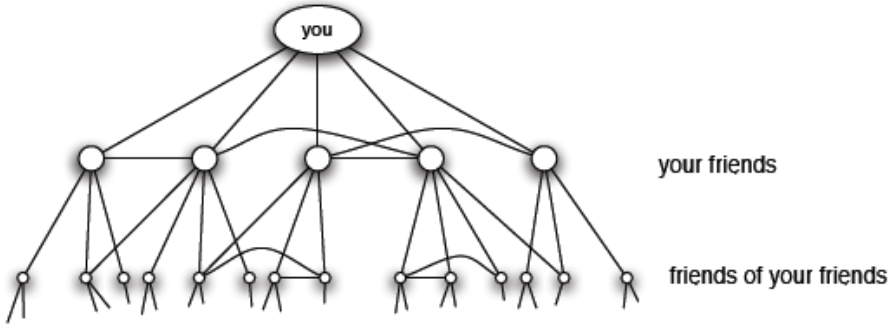
Two striking facts:

1. Short paths are abundant;
2. People are effective at collectively finding these short path;

The paradox of short paths abundance



(a) *Pure exponential growth produces a small world*



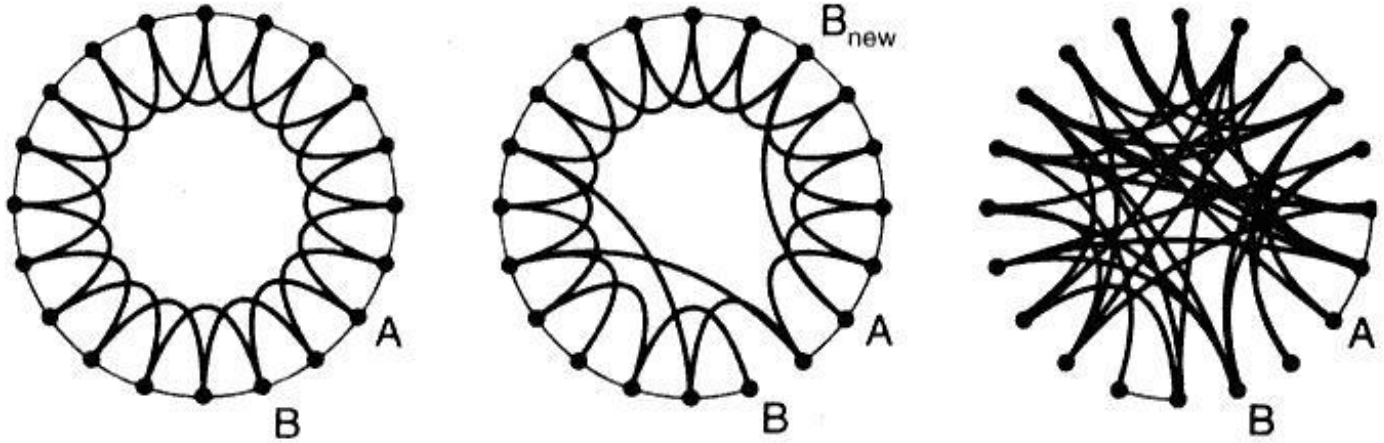
(b) *Triadic closure reduces the growth rate*

- Network grows exponentially, leading to the the existence of short paths!
  - The average person has between 500 and 1500 acquaintances, leading to $500^2 = 25K$ in one step, $500^3 = 125M$ in two steps, $500^4 = 62.5B$ in four (Figure (a)).
- However, the effect of *triadic closure* works to limit the number of people you can reach by following short paths (Figure (b)).
  - Triadic closure: If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.
- Question: Can we make up a simple model that exhibits both of the features: many closed triads (high clustering), but also very short path (small-world)?

Small world phenomenon:
Watts/Strogatz model

Reconciling two observations:

- **High clustering:** my friends' friends tend to be my friends
- **Short average paths**



The Watts-Strogatz small-world network

- Small-world network satisfies two properties according to Watts and Strogatz:
  - small average shortest path (global)
  - high clustering coefficient (local)
- Such a model follows naturally from a combination of two basic social-network ideas:
  - Homophily: the principle that we connect to others who are like ourselves, and hence creates many triangles.
  - Weak ties: the links to acquaintances that connect us to parts of the network that would otherwise be far away, and hence the kind of widely branching structure that reaches many nodes in a few steps.
- The crux of the Watts-Strogatz model: introducing a tiny amount of randomness–in the form of long-range weak ties–is enough to make the world "small" with short paths between every pair of nodes.

Watts-Strogatz model:

Generating small world graphs

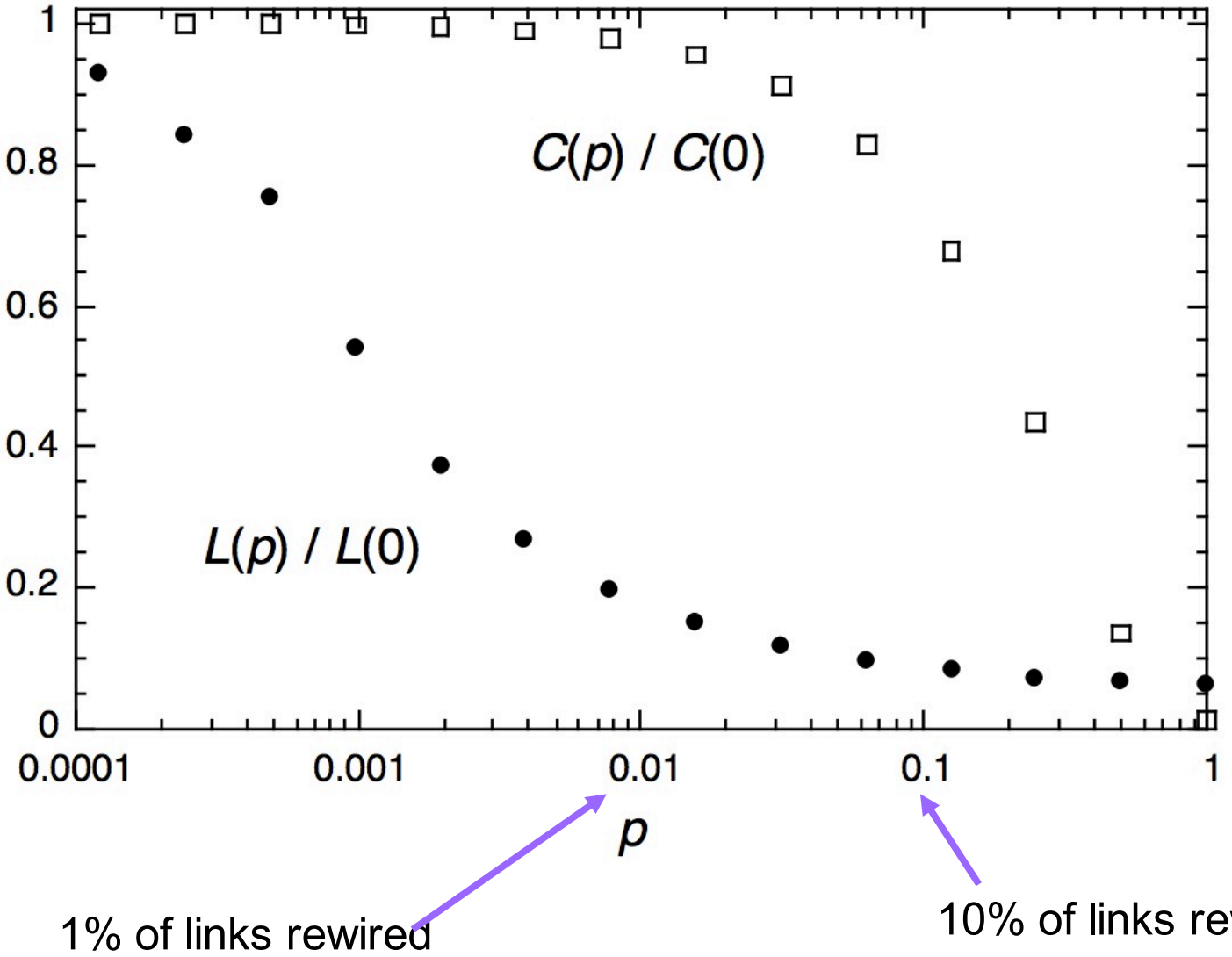
Select a fraction p of edges
Reposition on of their endpoints

Add a fraction p of additional edges leaving underlying lattice intact

- As in many network generating algorithms
  - Disallow self-edges
  - Disallow multiple edges

Clust coeff. and ASP as rewiring increases

- Fast decrease of average distance;
- Slow decrease in clustering (it remains almost constant, indicating that a transition to a small world is almost undetectable at a local level for p < 0.1



Hypothesis Testing with Network Data

Units of Analysis

- Dyadic (tie-level)
  - The raw data
  - Cases are pairs of actors
  - Variables are attributes of the relationship among pairs (e.g., strength of friendship; whether give advice to; hates)
  - Each variable is an actor-by-actor matrix of values by dyad
- Monadic (actor-level)
  - Cases are actors
  - Variables are aggregations that count number of ties a node has, or sum of distances to others (e.g., centrality)
  - Each variable is a vector of values, one for each actor
- Network (group-level)
  - Cases are whole groups of actors along with ties among them
  - Variables aggregations that count such things as number of ties in the network, average distance, extent of centralization, average centrality
  - Each variable has one value per network

Types of Hypotheses

- Dyadic (multiplexity)
  - Friendship ties lead to business ties
  - Social ties betweenm exchange partners leads to less formal contractual ties (embeddedness)
- Monadic
  - Actors with more ties are more successful (social capital)
- Mixed Dyadic-Monadic (autocorrelation)
  - People prefer to make friends (dyad level) with people of the same gender (actor level) (homophily)
  - Friends influence each other's opinions
- Network
  - Teams with greater density of communication ties perform better (group social capital)

Statistical Issues

- Samples non-random
- Often work with populations
- Observations not independent
- Distributions unknown
- This is not true if comparing network measures across independent networks
  - Then you can calculate the measures and input them to normal Regressoins
  - This is generally true in [pure] ego-net analysis

Solutions

- Non-independence
  - Model the non-independence explicitly as in Hierarchical LM
    - Assumes you know all sources of dependence
  - Permutation tests
- Non-random samples/populations
  - Permutation tests
- Unknown distributions
  - Permutation tests

Logic of Permutation Test

- Compute test statistic
  - e.g., correlation or difference in means
  - Correlation between centrality and salary is 0.384 or difference in mean centrality between the boys and the girls is 4.95.
  - Ask what are the chances of getting such a large correlation or such a large difference in means if the variables are actually completely independent?

- Wait! If the variables are independent, why would the correlation or difference in means be anything but zero?
  - Sampling
  - “Combinatorial chance”: if you flip coin 10 times, you expect 5 heads and 5 tails, but what you actually get could be quite different

Logic of Permutation Test

- So to evaluate an observed correlation between two variables of 0.384, we want to
  - correlate thousands of variables similar to the ones we are testing that we know are truly independent of each other, and
  - see how often these independent variables are correlated at a level as large as 0.384
    - The proportion of random correlations as large (or small) as the observed value is the p-value of the test
- How to obtain thousands of independent variables whose values are assigned independently of each other?
  - Fill them with random values
    - But need to match distribution of values
  - Permute values of one with respect to the other

Outline of Permutation Test

- Get observed test statistic
- Construct a distribution of test statistics under null hypothesis (no relationship)
  - Thousands of permutations of actual data
- Count proportion of statistics on permuted data that are as large as the observed
  - This is the p-value of the test

Friendship, age , class

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

≈

	A	B	C	D	E	F	G
A	0	1	0	2	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

+

	A	B	C	D	E	F	G
A	0	1	0	2	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

Friendship tie

Age difference

education

Friendship, age , class

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

≈

	A	B	C	D	E	F	G
A	0	1	0	2	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

+

	A	B	C	D	E	F	G
A	0	1	0	2	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

Friendship tie

Age difference

education

QAP procedure

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

≈

	A	B	C	D	E	F	G
A	0	1	0	2	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

+

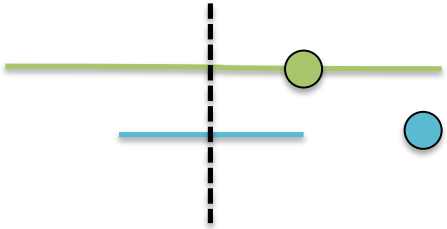
	A	B	C	D	E	F	G
A	0	1	0	2	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

Friendship tie

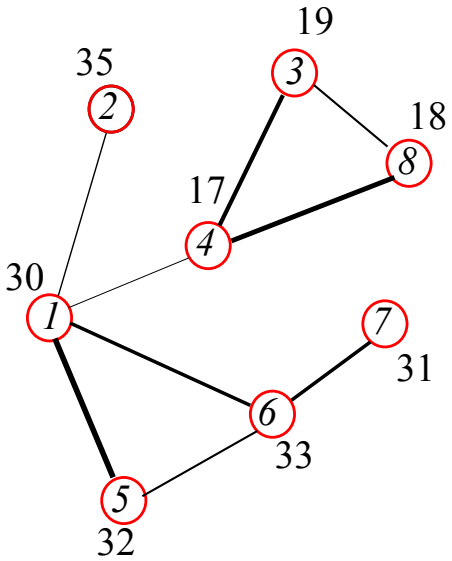
Age difference

education

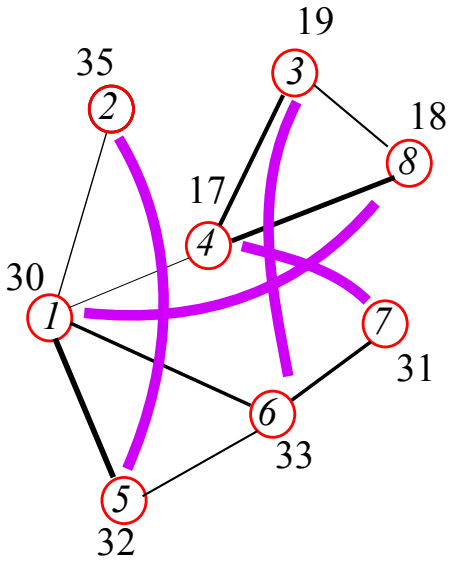
- Permutes dependent variables lots of time. Measure the sampling distribution of the coefficients.
- P-value is a proportion of times that the observation is Falling outside the sampling distribution.



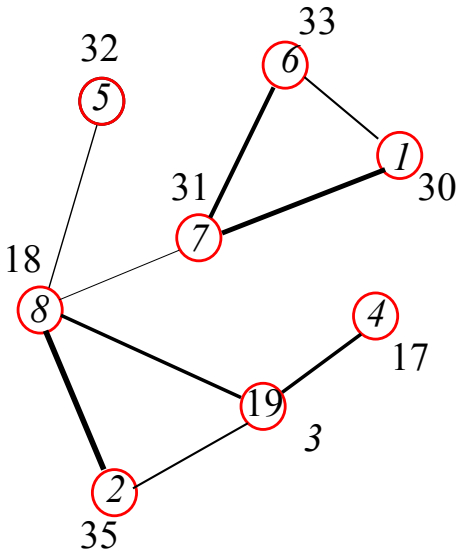
QAP process – graph representation



before



reshuffling



after

1. **Regression** on response and predictors;

2. **Permute response variable** lots of time to create random datasets
   a. gives sampling distribution of null hypothesis)
   b. Preserves dependence between dyads – (person A's values stay together during permutation)
   c. but removes relationship between response/predictor

Monadic Hypotheses

	Centrality	Grades
bill	10	2.1
maria	20	9.5
mikko	40	7.3
esteban	30	4.1
jean	70	8.1
ulrik	50	8.1
joao	40	6.6
myeong-gu	50	3.3
akiro	60	9.1
chelsea	10	7.2

- This, effectively, is basic social science research
  - However, centrality measures in most network based research are non-independent, so OLS is not appropriate
  - Ego-Net based research, on the other hand, would arguably yield independent measures

Testing Monadic Hypotheses

- We use the same techniques for determining coefficients as in traditional statistics
  - Regression for continuous variables
  - T-Tests to compare across two groups
  - ANOVA to compare across more than two
- But, we use the permutation test mechanisms to determine the significance of our findings

Dyadic Hypotheses

- Hubert / Mantel QAP test
  - All variables are actor-by-actor matrices
  - We use one relation (dyadic variable) to predict another
  - Test statistic is $\gamma = \sum_i \sum_j x_{ij} y_{ij}$
  - Significance is

$$prop(\gamma \geq \gamma^P),$$

$$\gamma^P = \sum_i \sum_j x_{ij} y_{p(i)p(j)}$$

- QAP correlation & MR-QAP multiple regression

Friendship

	Jim	Jill	Jen	Joe
Jim	-	1	0	1
Jill	1	-	1	0
Jen	0	1	-	1
Joe	1	0	1	-

} X

Proximity

	Jim	Jill	Jen	Joe
Jim	-	3	9	2
Jill	3	-	1	15
Jen	9	1	-	3
Joe	2	15	3	-

} Y

Dyadic/Monadic Hypotheses

- One dyadic (relational) variable, one monadic (actor attribute) variable
  - Technically known as autocorrelation
  - But, unlike in OLS, autocorrelation is **NOT** bad
- Diffusion
  - adjacency leads to similarity in actor attribute
    - Spread of information; diseases
- Selection
  - similarity leads to adjacency
    - Homophily: birds of feather flocking together
    - Heterophily: disassortative mating

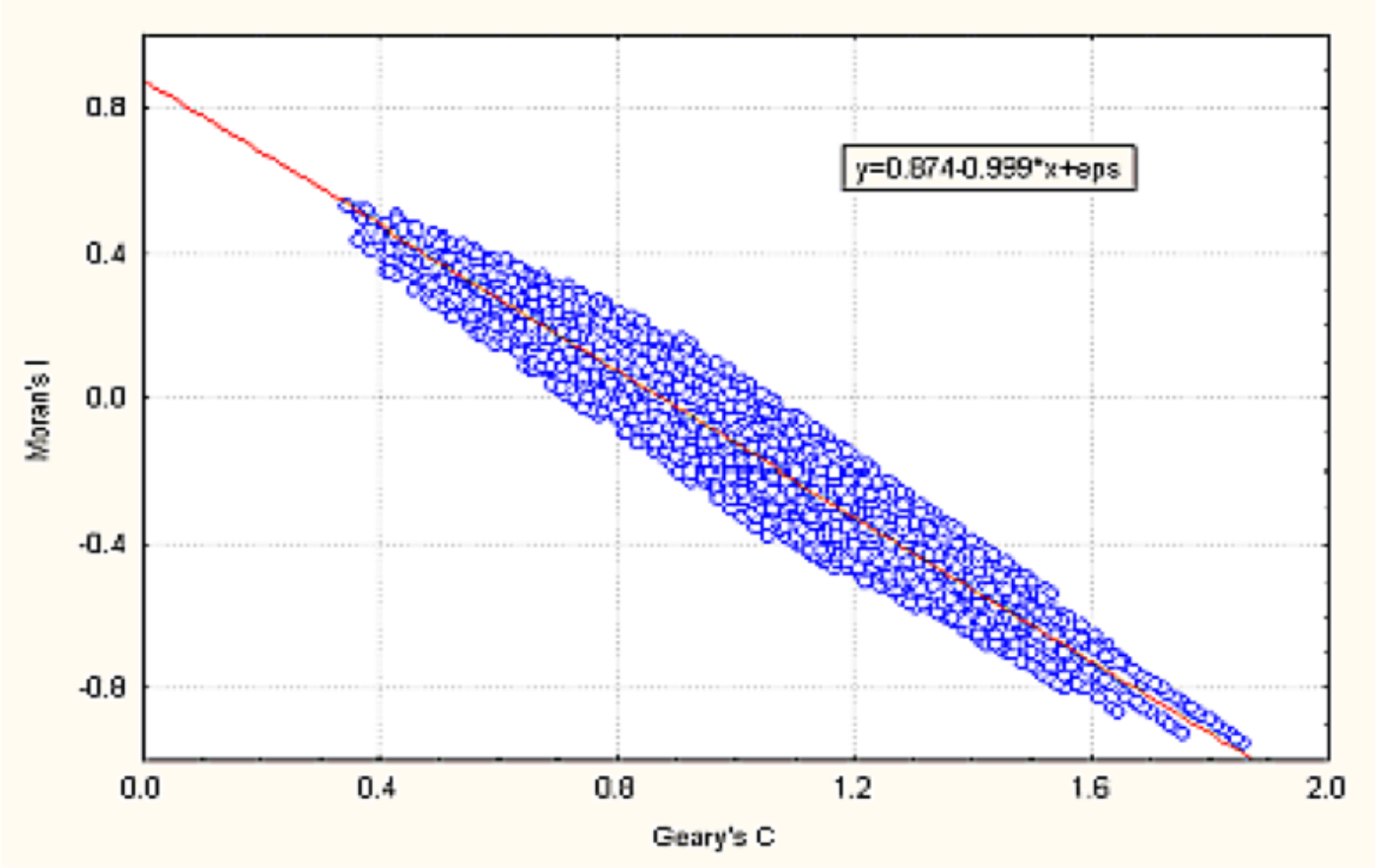
Continuous Autocorrelation

- Each node has score on continuous variable, such as age or rank
- Positive autocorrelation exists when nodes of similar age tend to be adjacent
  - Friendships tend to be homophilous wrt age
  - Mentoring tends to be heterophilous wrt age
- Can measure similarity via difference or product

Autocorrelation Measures

- Geary's C
  - Also called Geary's [Contiguity] Ratio
  - Most sensitive to local autocorrelation
- Moran's I
  - Measures autocorrelation not only on variable values or location (adjacency), but rather on both simultaneously
  - More sensitive to global autocorrelatoin
- I is about covariation of pairs, C is about variation in variable values
- Really the differences are probably immaterial

Comparing C & I



This figure suggests a linear relation between Moran's *I* and Geary's *C*, and either statistic will essentially capture the same aspects of spatial autocorrelation.

http://www.lpc.uottawa.ca/publications/moransi/moran.htm

Geary's C

- Let $w_{ij} > 0$ indicate adjacency of nodes i and j, and X_i indicate the score of node i on attribute X (e.g., age)

$$C = (n-1)\frac{\sum_i \sum_i w_{ij}(x_i - x_j)^2}{2\sum_{i,j} w_{ij} \sum_i (x_i - \bar{x})^2}$$

- Range of values: 0 <= C <= 2
  - C=1 indicates independence;
  - C > 1 indicates negative autocorrelation;
  - C < 1 indicates positive autocorrelation (homophily)

Moran's I

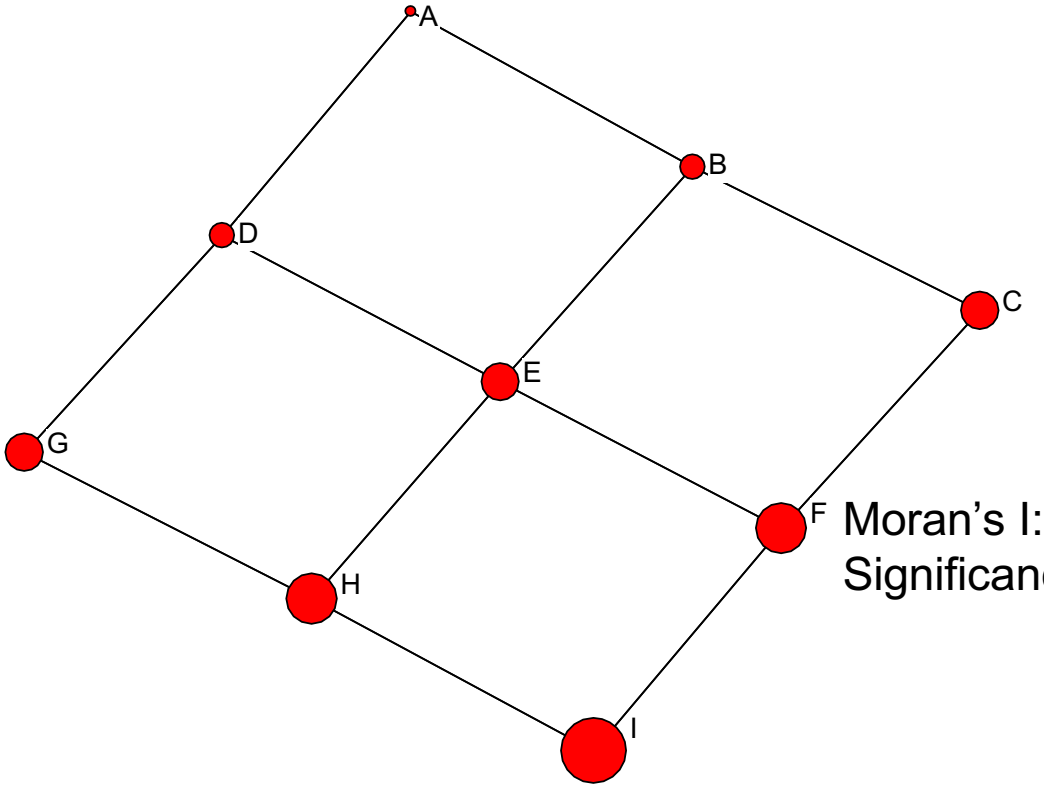
- Ranges between -1 and +1
- Expected value under independence is -1/(n-1)
- I → +1 when positive autocorrelation
- I → -1 when negative autocorrelation

$$\sum w_{ij}(x_i - x)(x_j - x)$$

$$I = n \frac{\sum_{i,j} w_{ij} \sum_i (x_i - \bar{x})^2}{\sum_{i,j} w_{ij}}$$

Positive Autocorrelation

(Similars adjacent; Moran's I > -0.125)



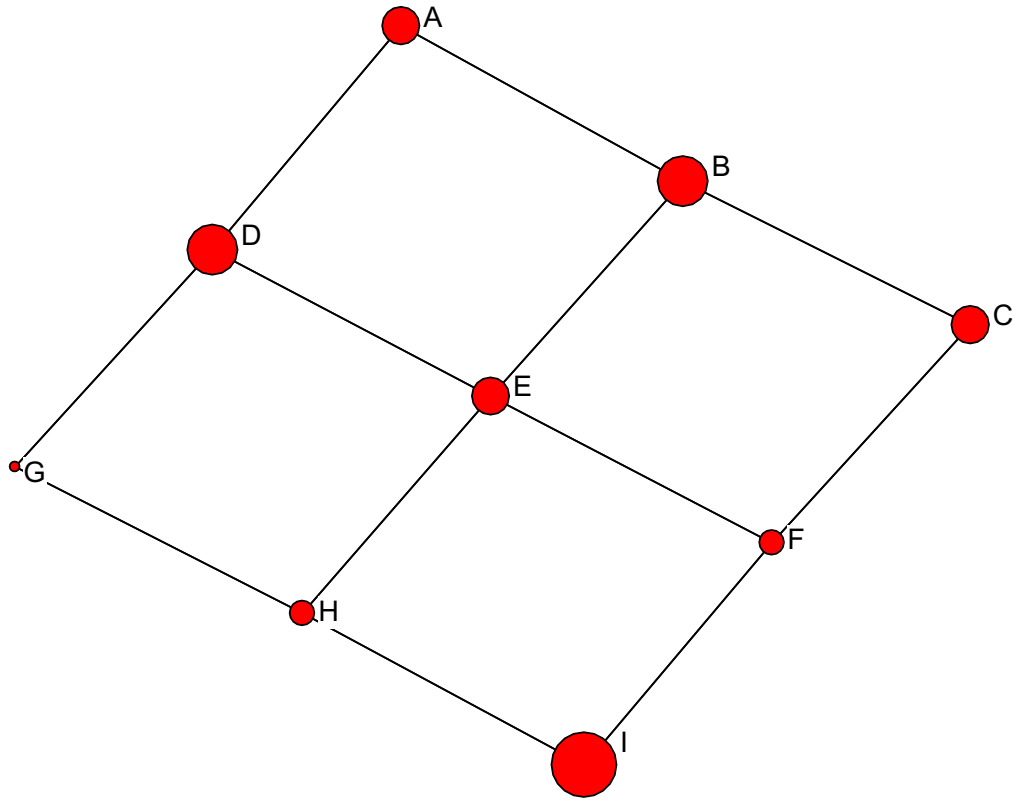
Node	Attrib A
A	1
B	2
C	3
D	2
E	3
F	4
G	3
H.	4
I.	5

Moran's I: 0.500

Significance: 0.000

No Autocorrelation

Independence; (Moran's I ≈ -0.125)

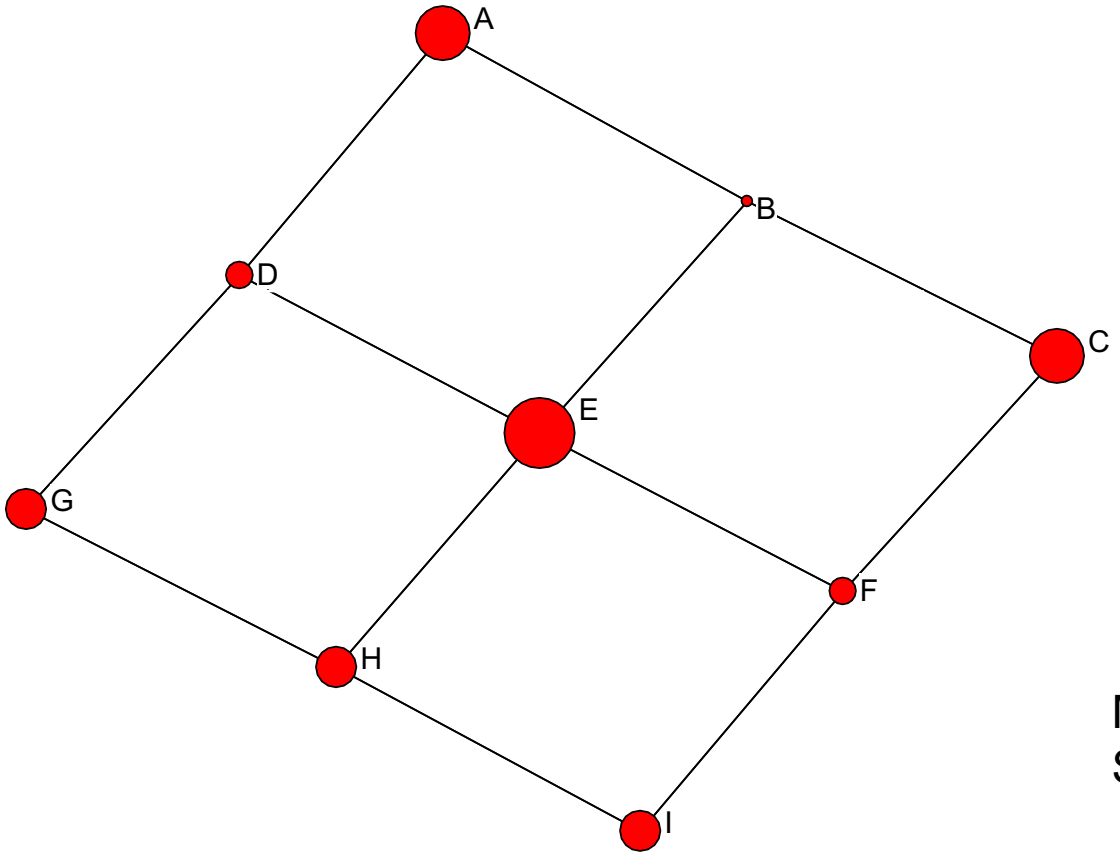


Node	Attrib
A	3
B	4
C	3
D	4
E	3
F	2
G	1
H	2
I	5

Moran's I: -0.250
Significance: 0.335

Negative Autocorrelation

(Dissimilars adjacent; Moran's I < -0.125)



Node	Attrib
A	4
B	1
C	4
D	2
E	5
F	2
G	3
H	3
I	3

Moran's I: -0.875
Significance: 0.000

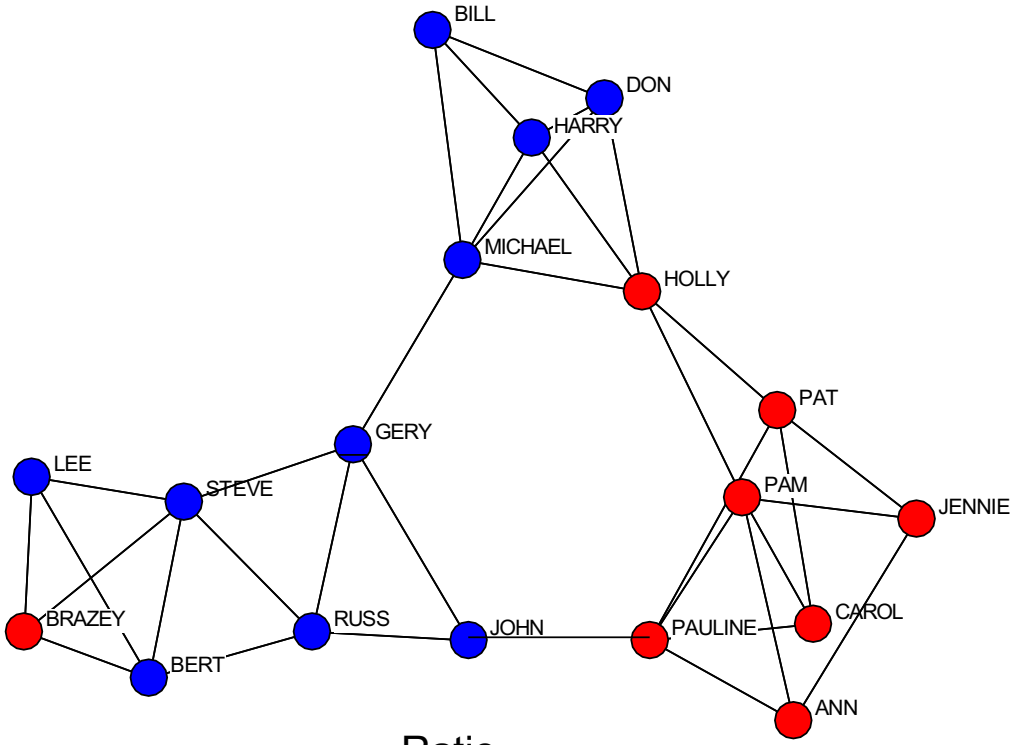
Interpreting Autocorrelation

- With Moran's *I*
  - A value near +1.0 indicates clustering (adjacency tends to accompany similarity along a dimension)
  - A value near -1.0 indicates dispersion (adjacency tends to accompany dissimilarity along a dimension)
  - a value near 0 indicates random distribution
- For Geary's C
  - just substitute 0, 2, and 1 for 1, -1, and 0 above

With Categorical Variables

- Moran's I and Geary's C are designed for continuous variables (also, frequently, dichotomous)
- For categorical variables, we use either ANOVA Density Models to determine if there is a homophily effect
- Homophily effects (preference for in-group ties) can be modeled as
  - Constant: Determine one in-group effect across all groups
    - People in general prefer their own gender to same extent, independent of their gender.
  - Variable: Each group can have its own in-group effect
    - Some groups show stronger tendencies to choose in-group ties than others.
    - E.g., Mormans show stronger in-group marriage ties than other Christian denominations

Campnet Example



Observed

	Female	Male
Female	12	7
Male	7	16

Ratio

	Female	Male
Female	1.87	0.38
Male	0.38	1.55

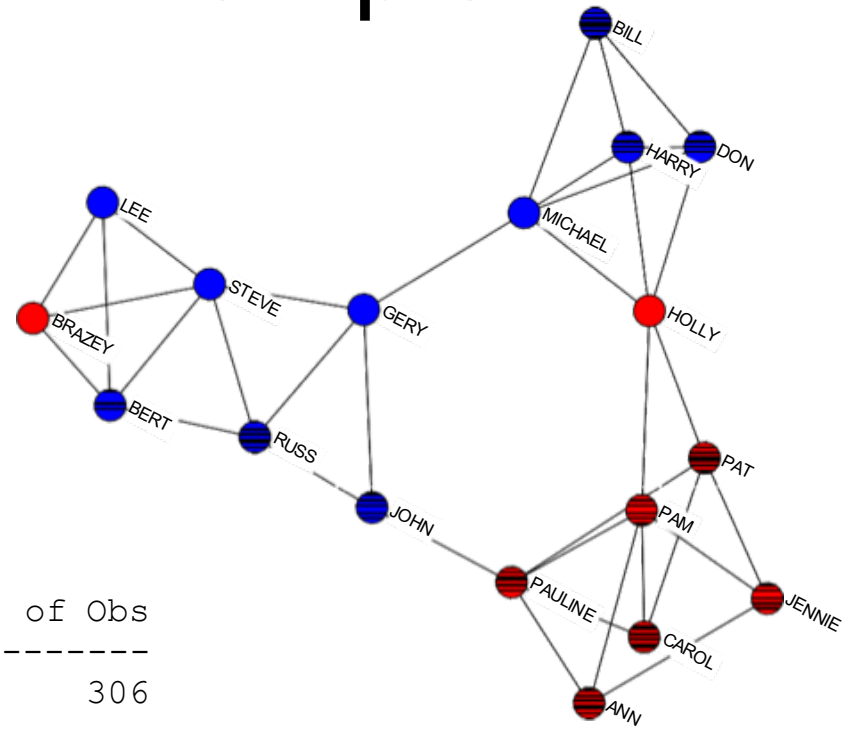
Expected

	Female	Male
Female	6.4	18.3
Male	18.3	10.3

Campnet Example

Density Table

	1 Femal	2 Male
1 Fem	0.429	0.087
2 Mal	0.087	0.356



MODEL FIT

R-square	Adj R-Sqr	Probability	# of Obs
0.127	0.124	0.001	306

REGRESSION COEFFICIENTS

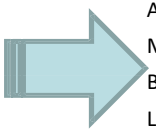
Independent	Un-stdized Coefficient	Stdized Coefficient	Significance	Proportion As Large	Proportion As Small
Intercept	0.087500	0.000000	1.000	1.000	0.001
Group 1	0.341071	0.313982	0.001	0.001	0.999
Group 2	0.268056	0.290782	0.001	0.001	0.999

Another Approach

- Convert the attribute vector into a matrix
- QAP this new matrix against the adjacency matrix
  - Significances will be the ~same because it uses same underlying permutation method
  - Values will follow same pattern (but not same values) as Moran's I

Using QAP for Autocorrelation

	Gender
HOLLY	1
BRAZEY	1
CAROL	1
PAM	1
PAT	1
JENNIE	1
PAULINE	1
ANN	1
MICHAEL	2
BILL	2
LEE	2
DON	2
JOHN	2
HARRY	2
GERY	2
STEVE	2
BERT	2
RUSS	2



	HOL	BRA	CAR	PAM	PAT	JEN	PAU	ANN	MIC	BIL	LEE	DON	JOH	HAR	GER	STE	BER	RUS
HOLLY	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
BRAZEY	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
CAROL	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
PAM	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
PAT	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
JENNIE	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
PAULINE	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
ANN	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
MICHAEL	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
BILL	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
LEE	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
DON	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
JOHN	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
HARRY	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
GERY	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
STEVE	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
BERT	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
RUSS	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

This matrix was constructed based on “exact match”
but you can use different transformations

Comparing QAP & Moran's I

Moran's I Output

A value of -0.059 indicates perfect independence.

Autocorrelation:	0.667
Significance:	0.001

QAP Output

Independent	Un-stdized Coefficient	Stdized Coefficient	Significance
Intercept	0.056250	0.000000	0.999
CAMPATTR2-MAT	0.251969	0.330131	0.001

A word about permutation test significances

- As you increase the number of iterations or permutations, the test statistic (correlation, difference in mean, etc.) will stay the same
- The p value, or significance, may change a little, but should converge
  - At relatively low permutations (2K), you may get different p values
  - A higher values (>25K or 50K) they should be stable and consistent

Inferential Network Analysis

ERGMS

A key twist on this simple model above is that while we work with dyads (i.e. our observations in the dataset will be ij dyads), the model is of the entire network – including all the dependencies.

Substantively, the approach is to *ask whether the graph in question is an element of the class of all random graphs with the given known elements.* For example, all graphs with 5 nodes and 3 edges, or, put probabilistically, the probability of observing the current graph given the conditions.

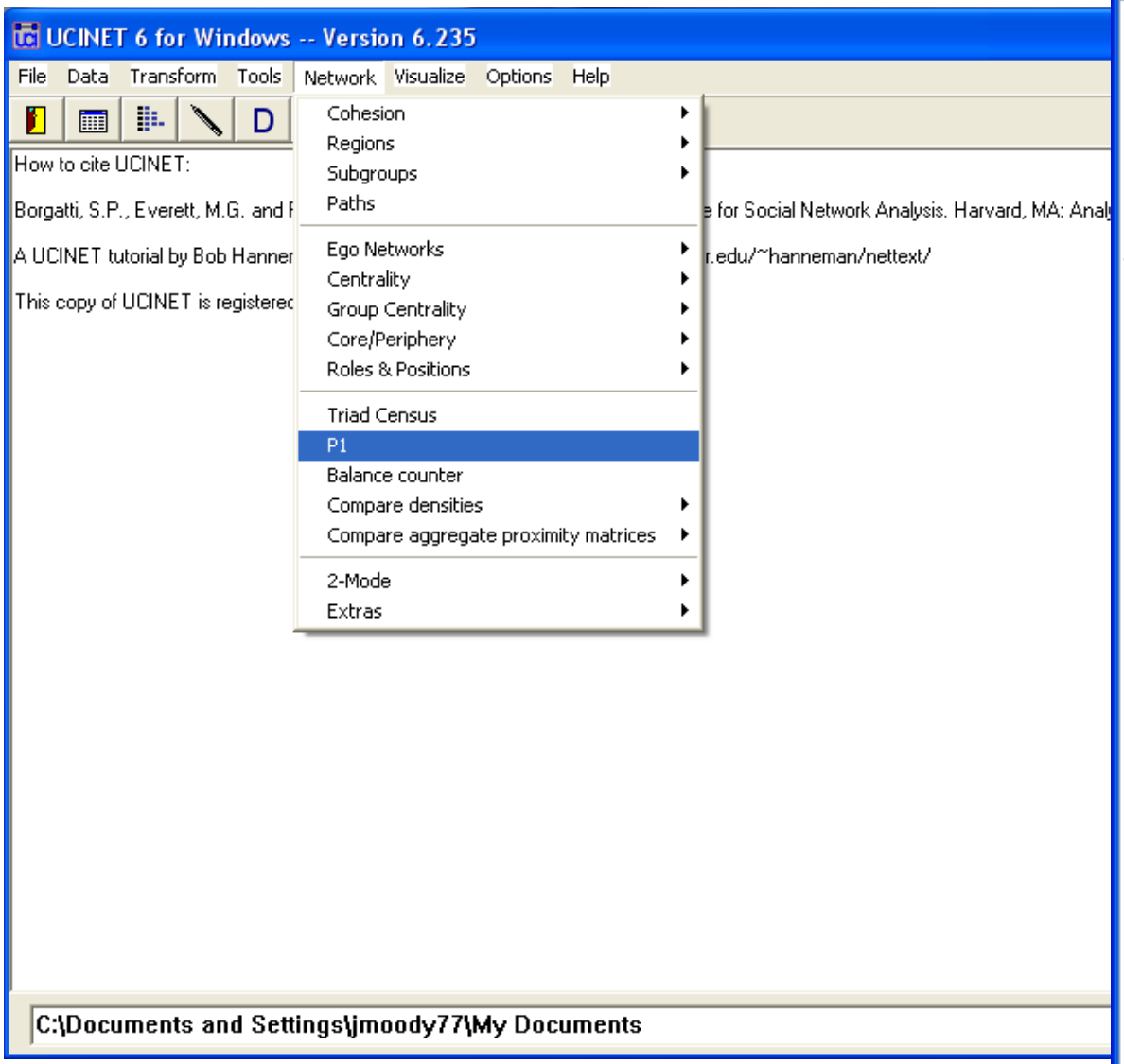
The “p1” model of Holland and Leinhardt is the classic foundation – the basic idea is that you can generate a statistical model of the network by predicting the counts of types of ties (asym, null, sym). They formulate a log-linear model for these counts; but the model is equivalent to a logit model on the dyads:

$$logit\left(X_{ij} = 1\right) = \alpha_i + \beta_j + \rho(X_{ij})$$

Note the subscripts! This implies a distinct parameter for every node *i* and *j* in the model, plus one for reciprocity.

Statistical Models for Networks

Modeling the network: ERGM



OUTPUT.LOG4 - Notepad

File Edit Format View Help

P1

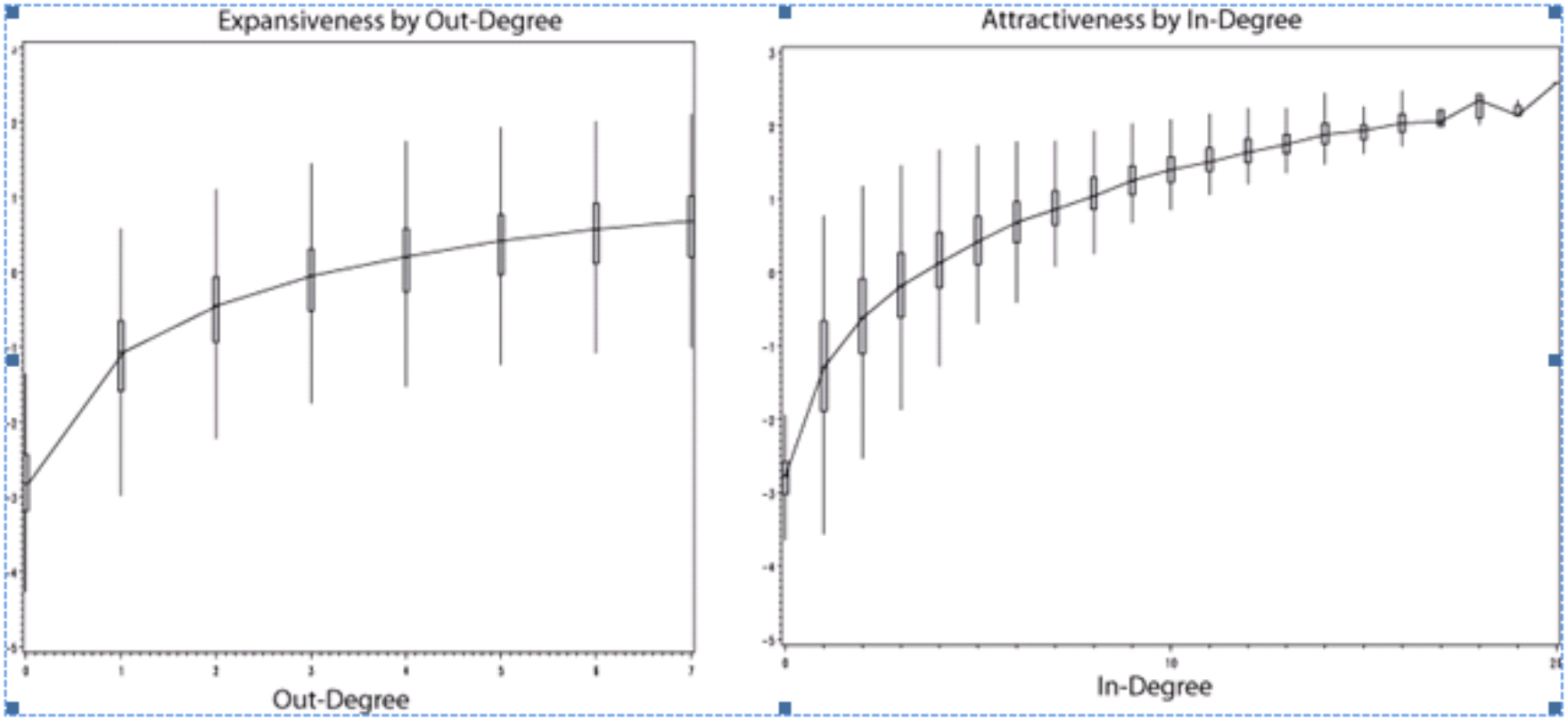
Input dataset: PRISON (C:\D

G-Square	DF
1166.61	4889

Theta = -4.1108
Rho = 4.4898

Expansiveness and Popularity Parameters

	1 Alpha	2 Beta
1	0.417	
2	0.602	-1.326
3	0.602	-1.326
4	-1.596	0.984
5	-0.148	0.539
6	-1.077	0.131
7	-1.596	0.984
8	1.303	-0.057
9	0.089	0.113
10	0.101	-1.042
11	-0.725	-0.690
12	0.298	0.317
13	-0.206	-0.176
14	0.101	-1.042
15	0.089	0.113
16	-1.272	1.547
17	-0.484	0.368
18	0.520	-0.124
19		0.056
20	0.089	0.113
21	-1.114	1.331
22	0.343	-0.454
23	0.343	-0.454
24	0.089	0.113
25		0.056
26		0.056
27	0.850	
28	-0.148	0.539



Results from SAS version on PROSPER datasets

Once you know the basic model format, you can imagine other specifications:

$$\text{Logit}(X_{ij} = 1) = \alpha_i + \beta_j + \rho(X_{ij})$$

$$\text{Logit}\,(X_{ij} = 1) = \alpha_i + \beta_j + \rho_g(X_{ij}) - \text{differential reciprocity}$$

$$\text{Logit}\,(X_{ij} = 1) = \alpha_i + \beta_j + \rho_g(X_{ij}) + \text{(node attributes)}$$

Key is to ensure that the specification doesn't imply a linear dependency of terms.

Model fit is hard to judge, and for all but the simplest rhs features, the se's are "approximate."

How to fix the inference problem?

Analytic & estimation solutions came with some careful thinking on the underlying structure on this model. Start with a re-expression of a general graph model:

$$p(X = x) = \frac{exp\{\theta' z(x)\}}{k(\theta)}$$

Where:

q is a vector of parameters (like regression coefficients)

z is a vector of network statistics, conditioning the graph

k is a normalizing constant, to ensure the probabilities sum to 1.

So here, we're just asking the probability of observing our network, given some network statistics.

We need a way to express the probability of the graph that doesn't depend on that constant. It turns out we can do this by conditioning on a 'complement' graph.

First some terms:

X_{ij}^{+}= Sociomatrix with ij element forced to be 1
X_{ij}^{-}= Sociomatrix with ij element forced to be 0
X_{ij}^{c}= Sociomatrix array without ij element

After some algebra:

$$log\left\{\frac{p(X_{ij}=1|X_{ij}^{c})}{p(X_{ij}=0|X_{ij}^{c})}\right\} = \theta'\left[z\left(x_{ij}^{+}\right) - z\left(x_{ij}^{-}\right)\right] = \theta'\delta(x)$$

We can re-write the probability of the graph as a function of the change scores (complement graph), which has to do with the tie being present or absent.

Which ends up being a logit model on z, where z are "change statistics" or counts of features on the full graph when that statistic for the ij dyad is differenced.

Now we can get an unbiased estimation of the graph as a function of the change statistics;

Imagine what the change score looks like for the simples configuration: an edge. This gives us an intercept only model: what's the number of ties in the network if each edge is/is not present?

What about reciprocity? What's the number of reciprocal ties if Xij is present/asbent.

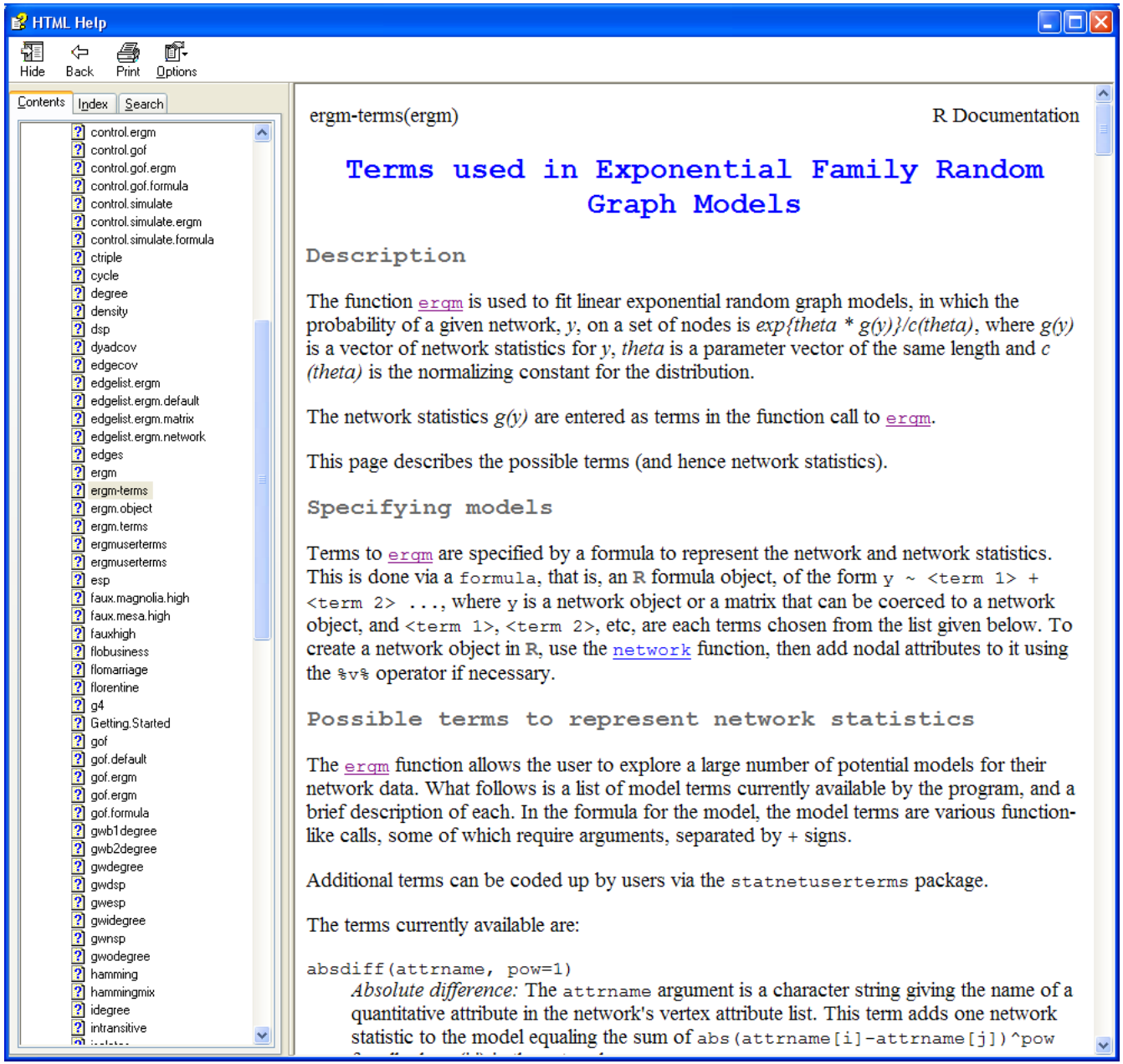
Steps in estimating an ERGM

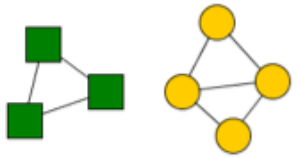
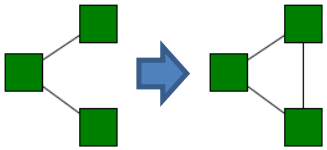
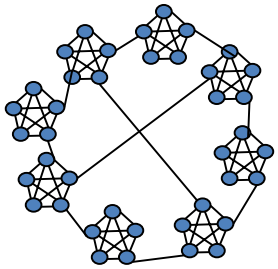
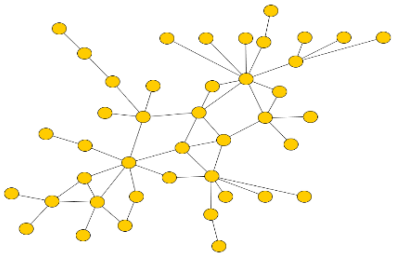
1) Specify the model
2) Fit the model
3) Examine MCMC chains for convergence & such
4) Examine Goodness of fit
    1) If poor, return to 1
    2) Else, publish your paper. ☺

Question is the likelihood of a network given an observed set of network mixing statistics.

The set of such statistics ("terms") is large…and growing.

Intuitively, these capture a social process you think is driving network formation.



Theory	Colloquialism	Structural Signature	Model Term
Homophily	*Birds of a feather…*		*NodeMatch()*
Social Balance	*A friend of a friend...* *A friend of an enemy…*		*Balance, Transitivity, GWESP*
Small-Worlds	*Don't I know your…* *or* *Kevin Bacon game…*		*Clustering & k-paths*
Preferential Attachment	*Rich get richer..* *First mover advantage*		*In-degree, k-stars*

Common classes of terms:

Term	Why?
Edges	Density
Receiver, Sender	Fit person specific degree distribution
Degree(d,attr)	Fit the observed global degree distribution, perhaps by attribute
Mutuality	Reciprocity
Nodecov(attr), nodefactor()	Differential row/colloumn effects by an attribute
Nodematch(attr)	Homophily on a particular attribute
Gwesp	Geometric form for closed partners
Dyadcov, edgecov	Pair specific covariates, differ by directed or not.
Isolates	Fit the number of isolated nodes in the graph
Cycle(k)	Fit cycles of length k (slow!)

Model Sensitivity

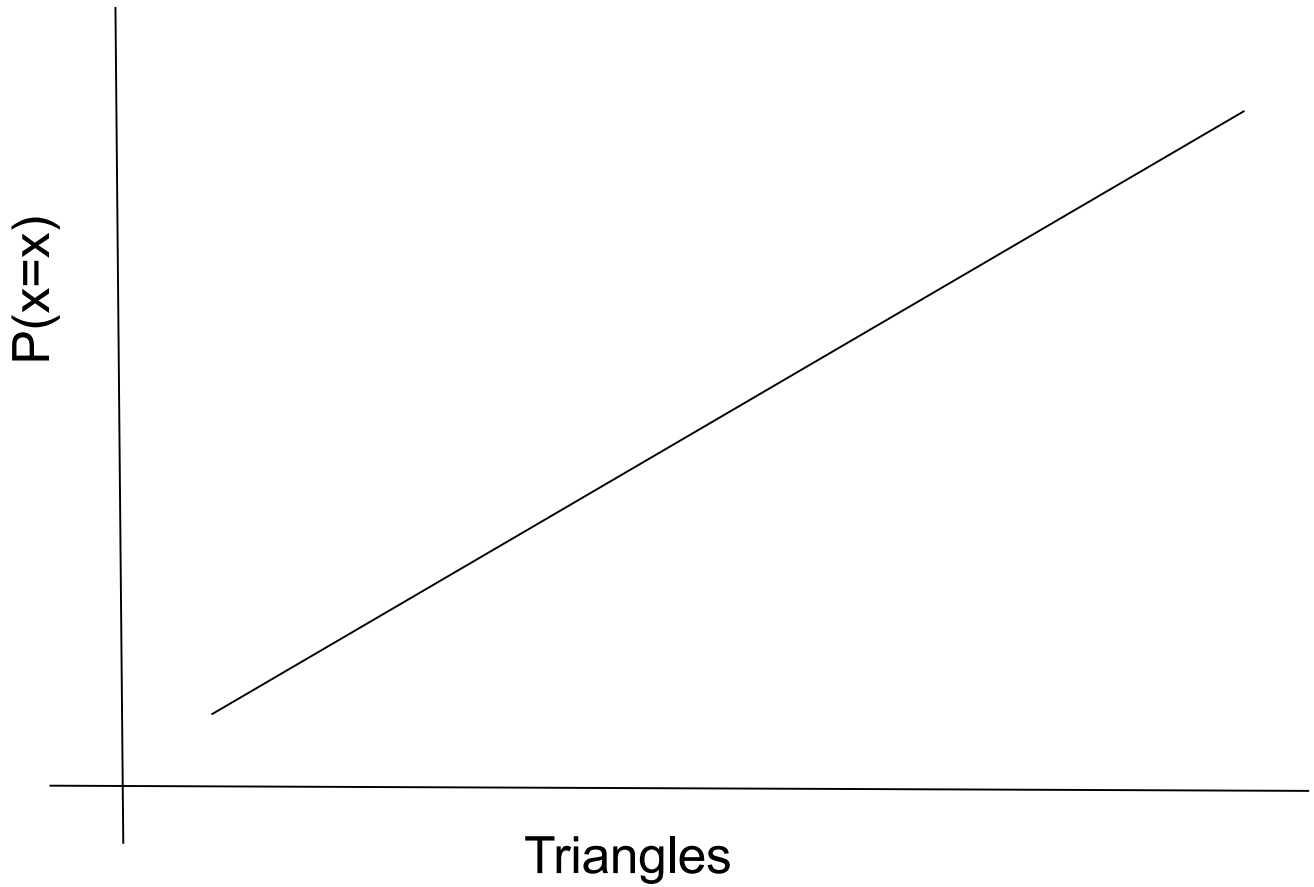
ERGM models are very sensitive to model specification, and work best if you have a good intuition about how the interdependencies in a network operate – **most of us do not have that intuition!**

Model Degeneracy: Intuitively, it happens when the network sample space implied by the model does not contain any instances of your model.

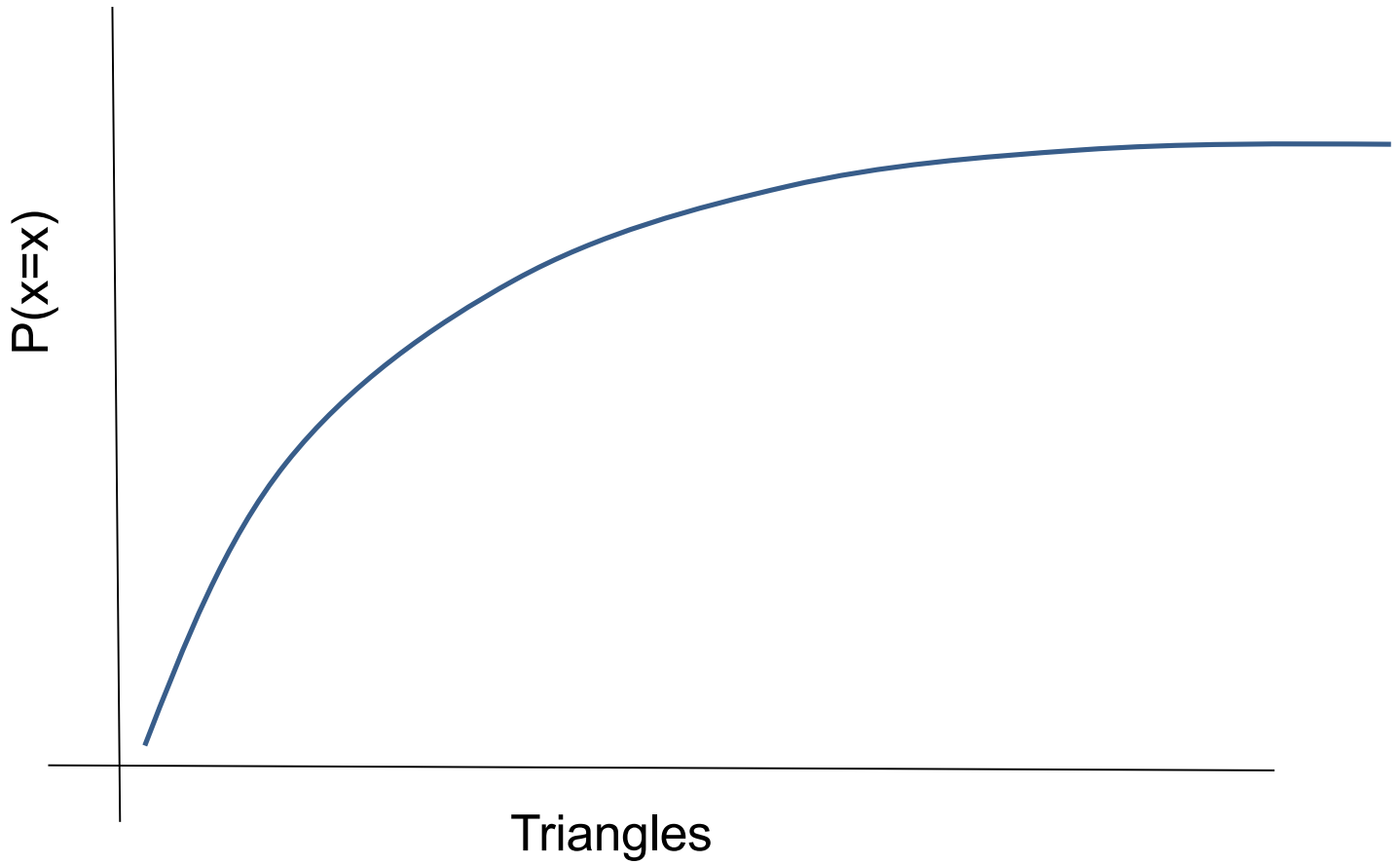
Example: Simple model of edges & triangles.

Intuitively, we'd expect from balance a positive coefficient on triangles.

Intuition from regression: b(triangle) is positive



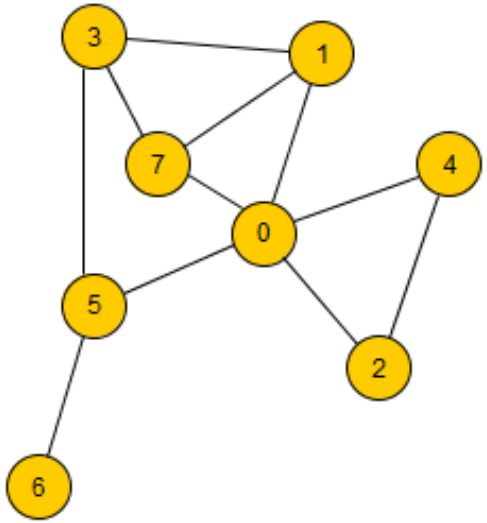
..so what you really want is:

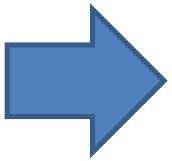


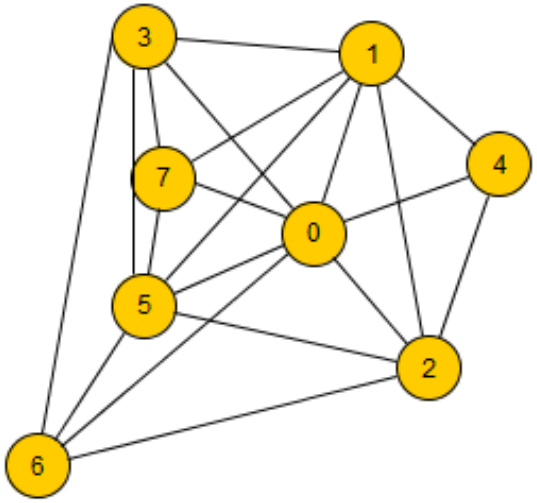
Or that there are marginal decreasing returns to each *additional* closed triad

GWESP

But note the model really says “more closed triads is good”







So if this is good…

..this is better!

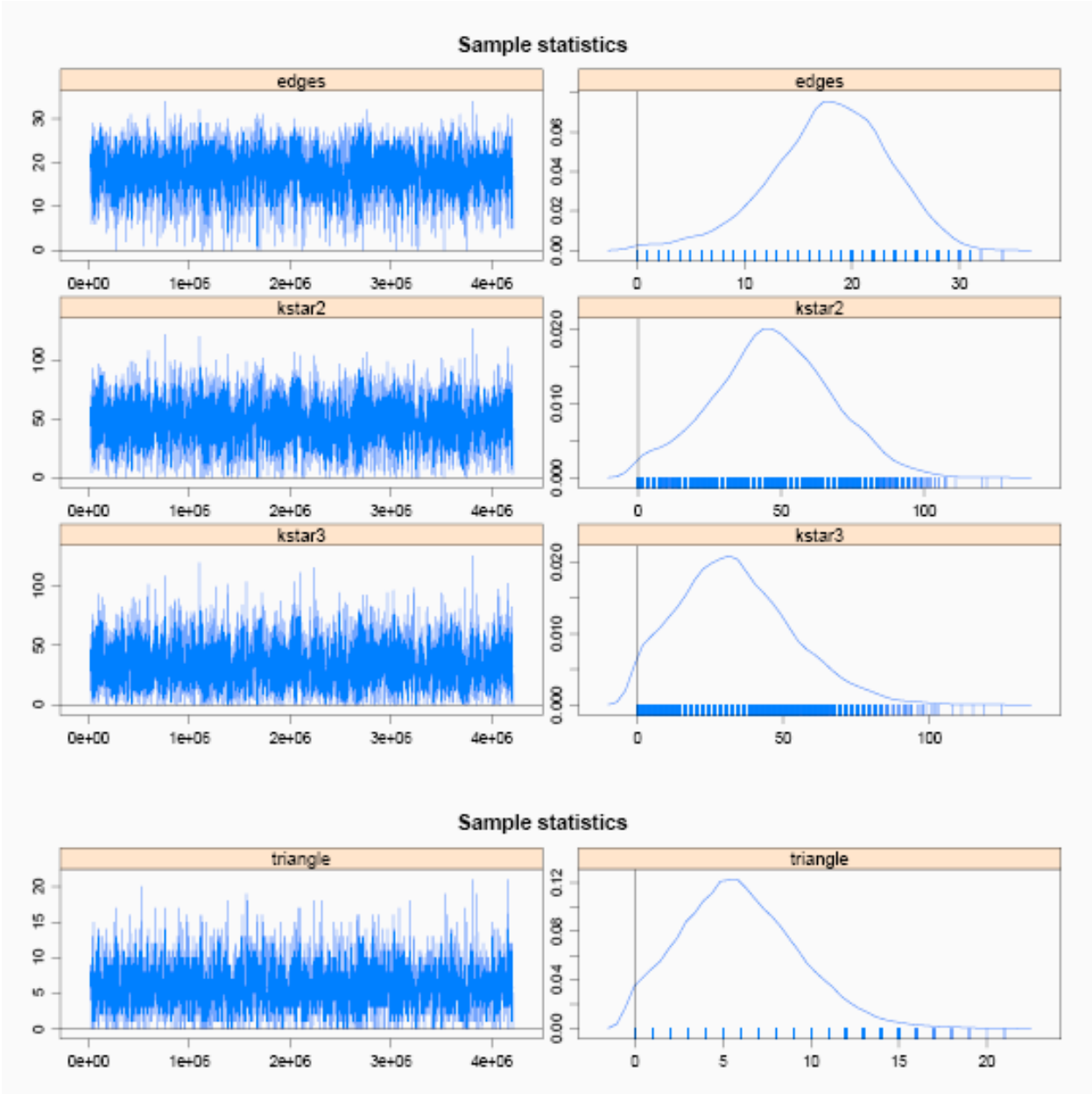
Running a model feels a lot like any general linear model:

```
R> model2 <- ergm(fmh ~ edges + nodematch("Grade") + nodematch("Race") +
+    nodematch("Sex"))
R> summary(model2)

...
                Estimate Std. Error MCMC s.e. p-value
edges          -10.01277    0.11526        NA  <1e-04 ***
nodematch.Grade  3.23105    0.08788        NA  <1e-04 ***
nodematch.Race   1.19646    0.08147        NA  <1e-04 ***
nodematch.Sex    0.88438    0.07057        NA  <1e-04 ***
```

Under the hood, it's using a pseudo-=likelihood (logit) for models with only dyad-independent features, or fitting an MCMC if there are dependencies.

Coeficcients are given in log-odds scale. If we exponentiate, we get the probability of observing a tie in the network

STATNET has a bunch of MCMC diagnostic tools. For example, you want to make sure your trace plots are nice and random, rather than trending in one direction or another…

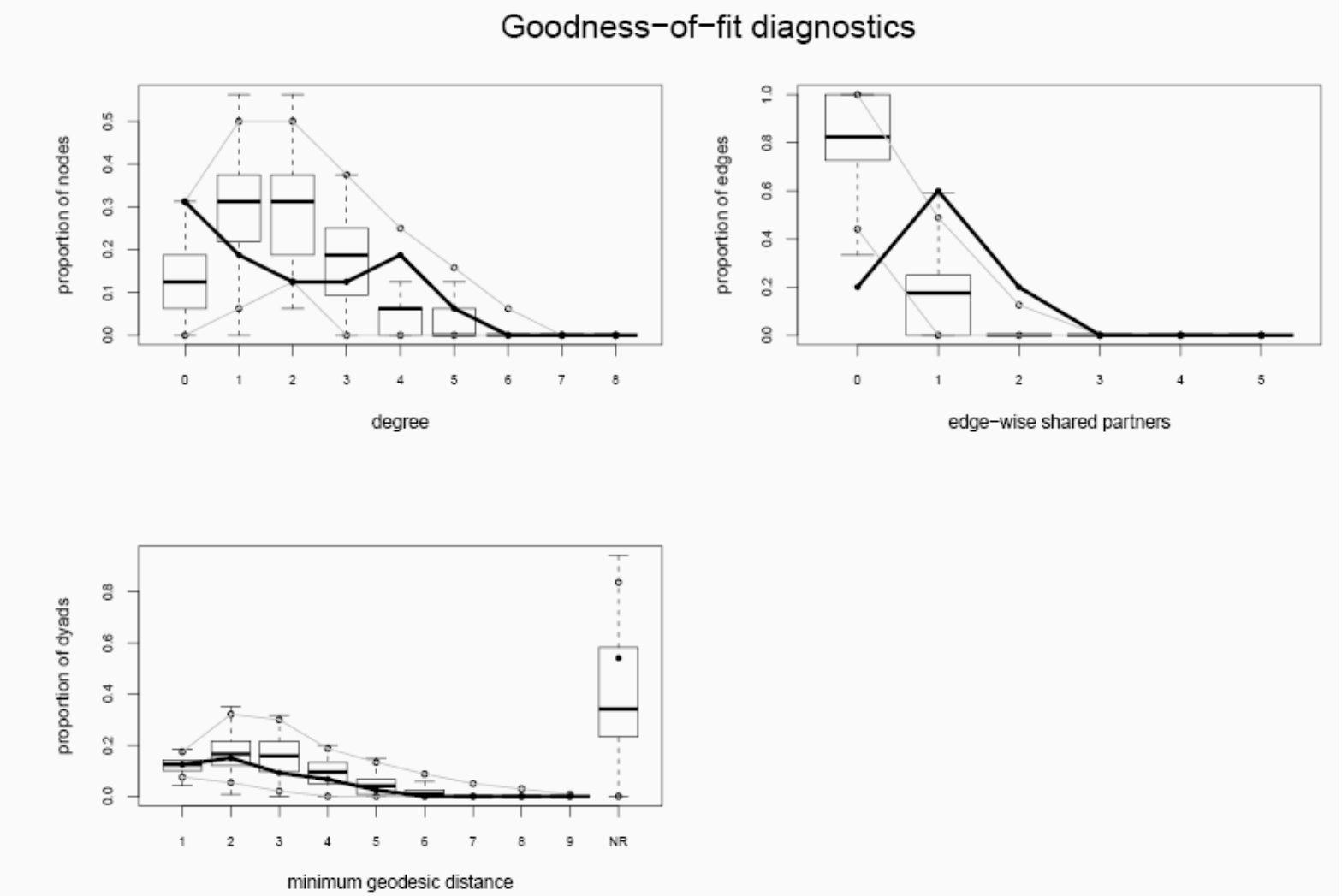


Once you have a model, the most common way to assess fit is to draw samples from the implied network space and compare them to your observed graph.

```
R> model2 <- ergm(fmh ~ edges + nodematch("Grade") + nodematch("Race") +
+    nodematch("Sex"))
R> summary(model2)

...
                Estimate Std. Error MCMC s.e. p-value
edges           -10.01277    0.11526        NA  <1e-04 ***
nodematch.Grade   3.23105    0.08788        NA  <1e-04 ***
nodematch.Race    1.19646    0.08147        NA  <1e-04 ***
nodematch.Sex     0.88438    0.07057        NA  <1e-04 ***
```

```
R> sim2 <- simulate(model2, burnin = 1e+6, verbose = TRUE, seed = 9)
```

Once you have a model, the most common way to assess fit is to draw samples from the implied network space and compare them to your observed graph.



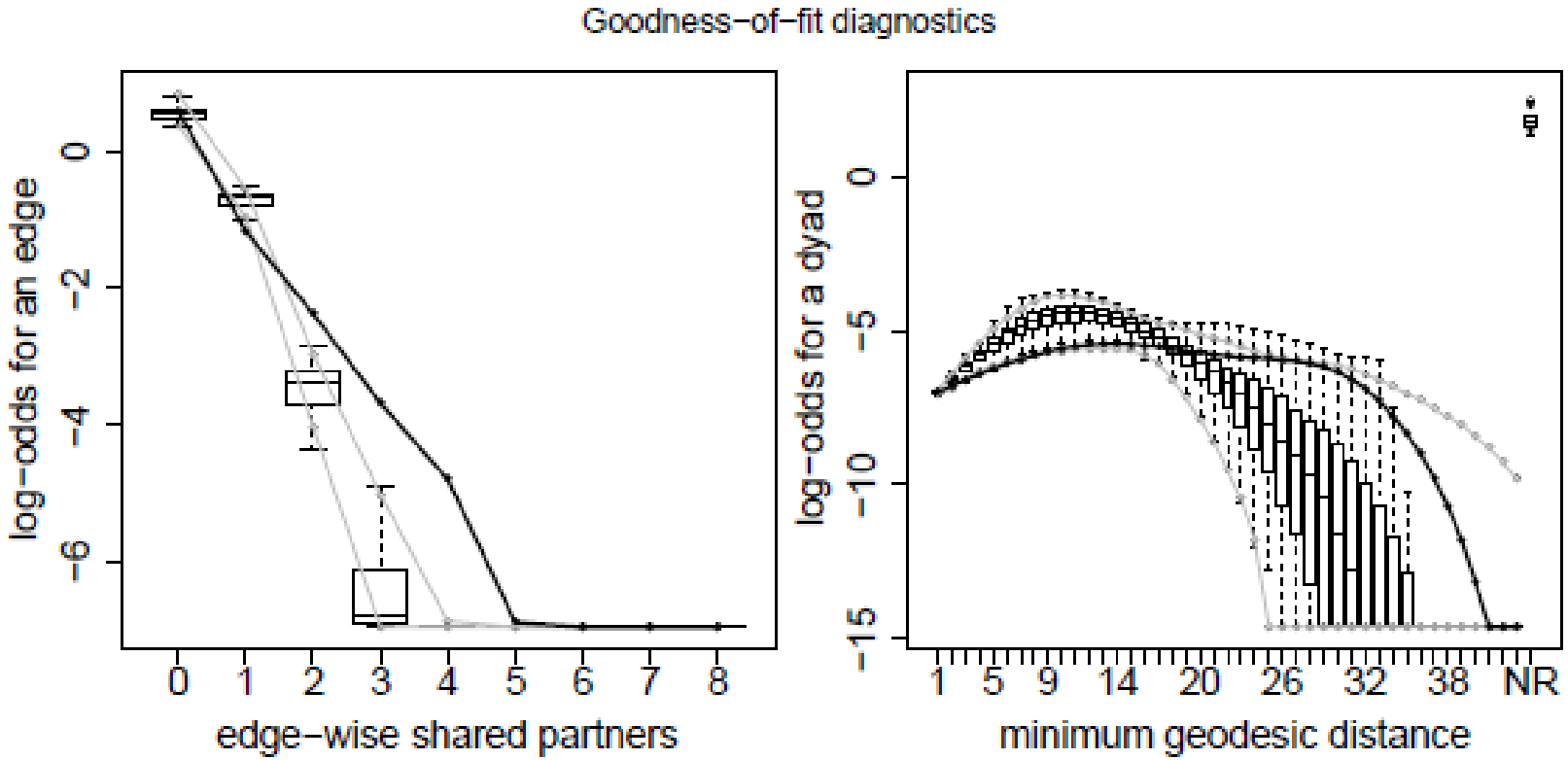
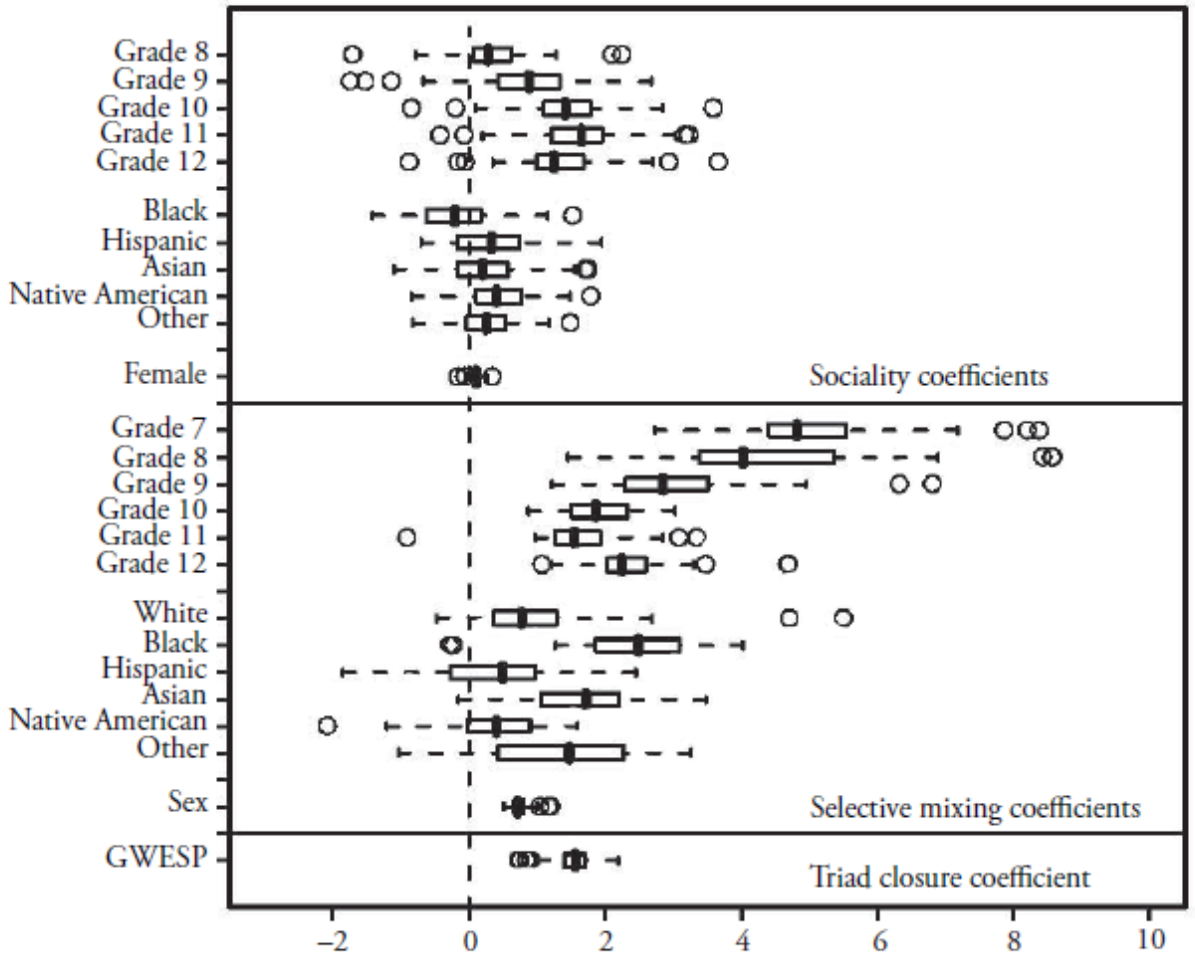

Goodness−of−fit diagnostics
log−odds for an edge
0
−2
−4
−6
0 1 2 3 4 5 6 7 8
edge−wise shared partners
log−odds for a dyad
0
−5
−10
−15
1 5 9 14 20 26 32 38 NR
minimum geodesic distance

Figure 3. Coefficients From the Full Model, Plotted Across All 59 Schools



Notes: Boxplots follow the Tukey method. Boxes represent quartiles; whiskers extend to the most extreme data point within 1.5 times the interquartile range from the edge of the box; and points represent outliers.

Lord of Flies theory is correct among adolescents:
We give them no structure and they create a rich hierarchy and beat the shit out of each other

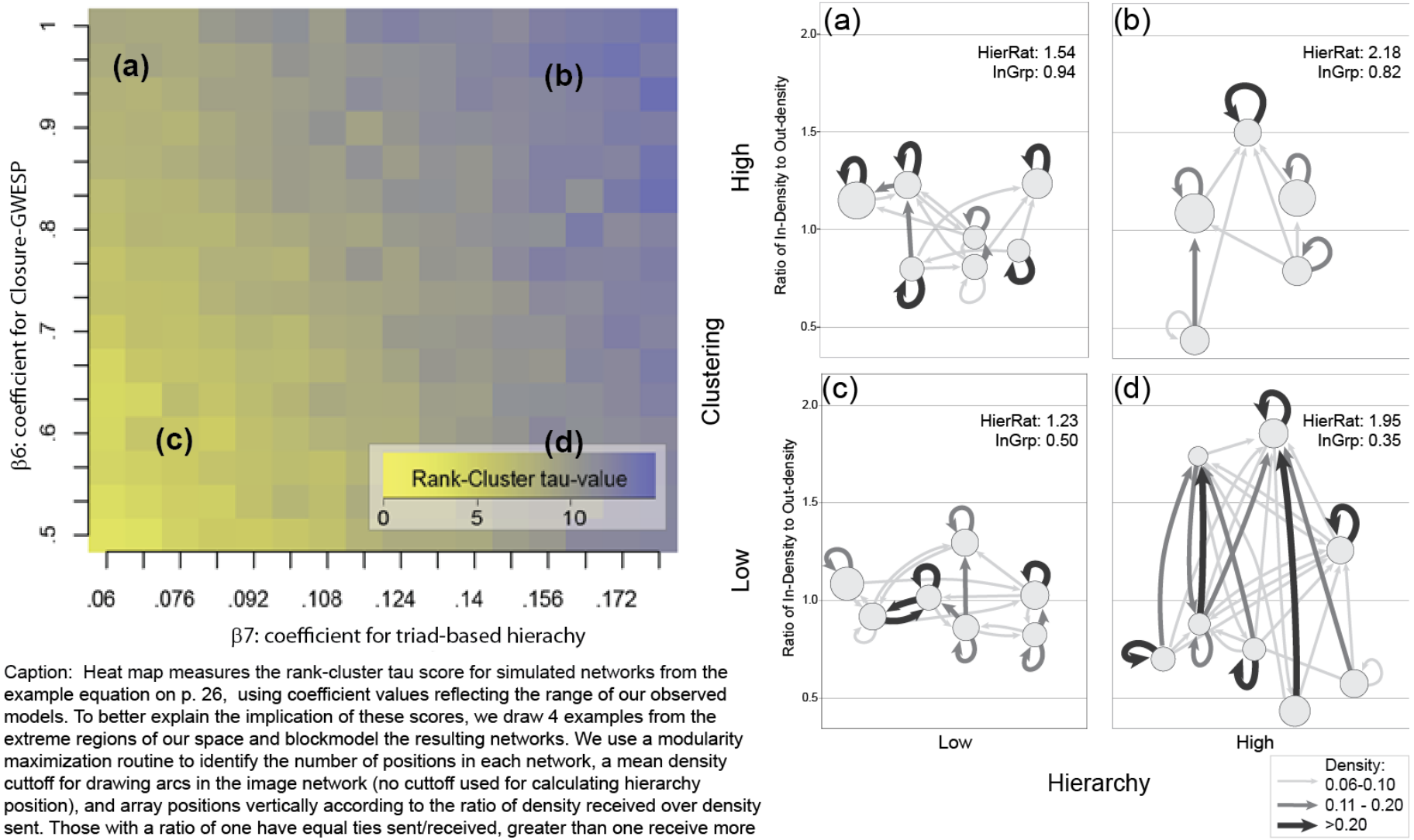
Network Ecology and Adolescent Social Structure

Daniel A. McFarland,[a] James Moody,[b,c] David Diehl,[d] Jeffrey A. Smith,[e] and Reuben J. Thomas[f]

Table 2. Mechanisms of Friendship Formation across Groups: Results of Multilevel Models with Measurement Error Correction

	School Networks							
	Cross Sectional				*Longitudinal*			
	Odds	Log-Odds	*t*		Odds	Log-Odds	*t*	
Edges	.001	–6.689	–128.88	***	.001	–6.615	–12.76	***
Mutuality	29.767	3.393	71.51	***	21.802	3.082	8.60	***
Closure	2.595	.954	111.07	***	2.372	.864	15.43	***
Hierarchy	1.131	.123	64.34	***	1.177	.163	11.54	***
Club Ties	1.508	.411	28.21	***	1.664	.509	5.68	***
Prior Year					26.552	3.279	13.12	***
Same Race	1.647	.499	20.05	***	3.221	1.170	6.96	***
Same Gender	1.202	.184	25.50	***	1.492	.400	7.38	***
Same Age	3.425	1.231	37.16	***	2.035	.710	7.22	***
GPA Diff.	.813	–.207	–35.13	***	.828	–.188	–5.56	***
SES Diff.	.966	–.035	–19.32	***	1.001	.001	.04	ns
Null Model Likelihood Ratio Test								
χ-square			1256.9	***			100.3	***
Model Fit								
Deviance			–1302.4				69.5	

Figure 1. Variability in hierarchical macro structure resulting from variability in micro-structural parameters with detailed block models from selected regions. (based on ERGM simulations from observed parameter estimate ranges)



Caption: Heat map measures the rank-cluster tau score for simulated networks from the example equation on p. 26, using coefficient values reflecting the range of our observed models. To better explain the implication of these scores, we draw 4 examples from the extreme regions of our space and blockmodel the resulting networks. We use a modularity maximization routine to identify the number of positions in each network, a mean density cuttoff for drawing arcs in the image network (no cuttoff used for calculating hierarchy position), and array positions vertically according to the ratio of density received over density sent. Those with a ratio of one have equal ties sent/received, greater than one receive more than they send, less than one send more than they receive.

Latent Space Models

Latent Space Approaches to Social Network Analysis

Peter D. HOFF, Adrian E. RAFTERY, and Mark S. HANDCOCK

Network models are widely used to represent relational information among interacting units. In studies of social networks, recent emphasis has been placed on random graph models where the nodes usually represent individual social actors and the edges represent the presence of a specified relation between actors. We develop a class of models where the probability of a relation between actors depends on the positions of individuals in an unobserved "social space." We make inference for the social space within maximum likelihood and Bayesian frameworks, and propose Markov chain Monte Carlo procedures for making inference on latent positions and the effects of observed covariates. We present analyses of three standard datasets from the social networks literature, and compare the method to an alternative stochastic blockmodeling approach. In addition to improving on model fit for these datasets, our method provides a visual and interpretable model-based spatial representation of social relationships and improves on existing methods by allowing the statistical uncertainty in the social space to be quantified and graphically represented.

KEY WORDS: Conditional independence model; Latent position model; Network data; Random graph; Visualization.

Fitting Position Latent Cluster Models for Social Networks with latentnet

Pavel N. Krivitsky
University of Washington

Mark S. Handcock
University of Washington

Does not require any theoretical machinery about social processes.

Simple latent distance model, where the z are actors positions in a latent space, such that people close to each other in z space tend to have a tie, and not otherwise:

$$P(Y|\alpha, Z) = \prod_{i \neq j}^{n} p(y_{i,j}|\alpha, z_i, z_j)$$

and

$$\text{logit } p(y_{i,j} = 1|\alpha, z_i, z_j) = \alpha - |z_i - z_j|,$$

Given a distribution of points in the space defined by z, probability of a tie decreases with their distance in the latent space.

Z can be as many dimensions as you want; typically we try to fit the minimum number of dimensions that provide reasonable fit to the data.

We don’t know what z means!

2d solution for Sampson monistary data

Don't require social processes or functional forms.

Works well, people close in z space have a tie

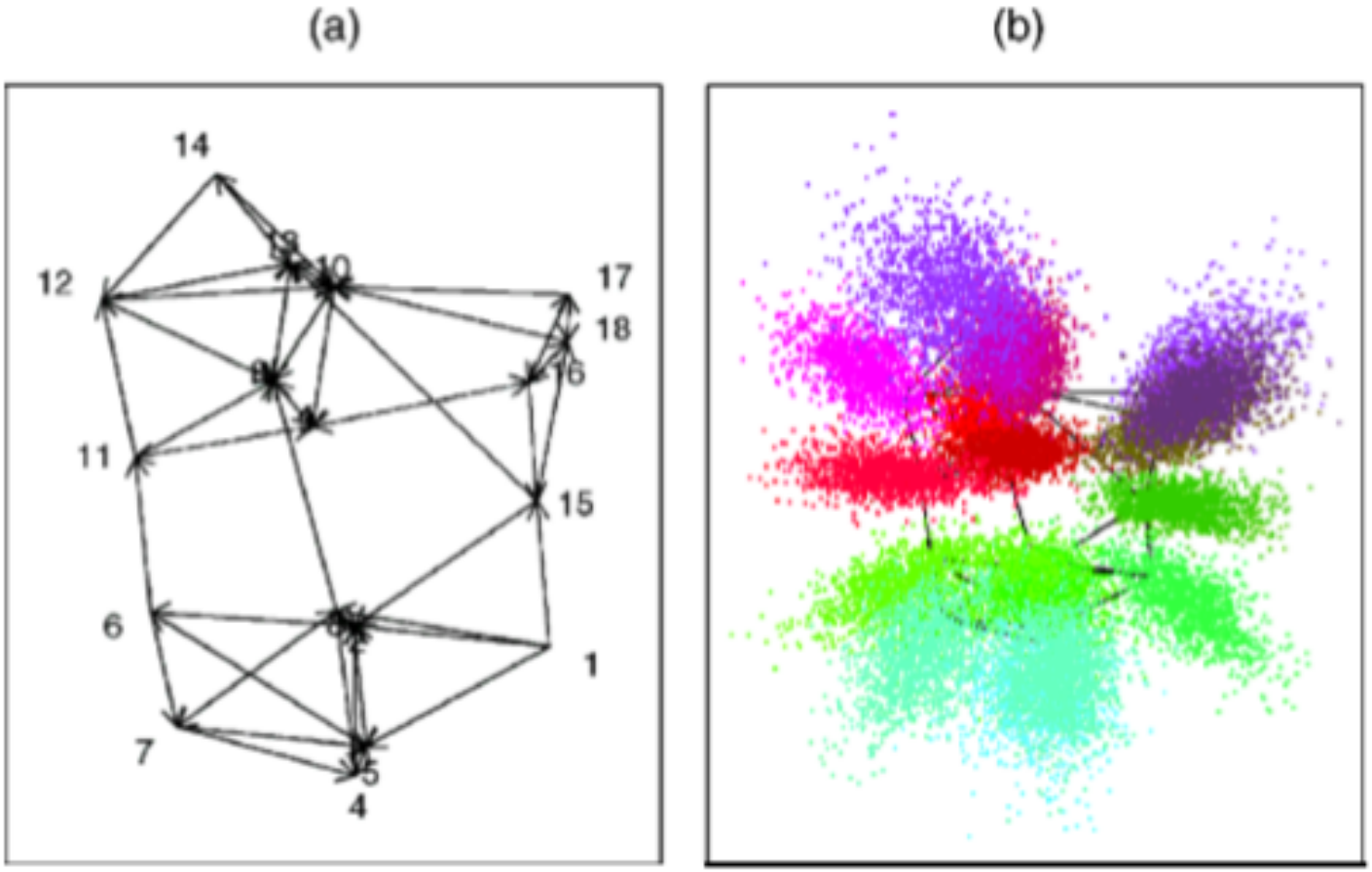


Figure 1. Maximum Likelihood Estimates (a) and Bayesian Marginal Posterior Distributions (b) for Monk Positions. The direction of a relation is indicated by an arrow.

Z = a dimension in some unknown space that, once accounted for makes ties independent.

In addition, we can now embed z within a group structure, which adds probability of ingroup ties.

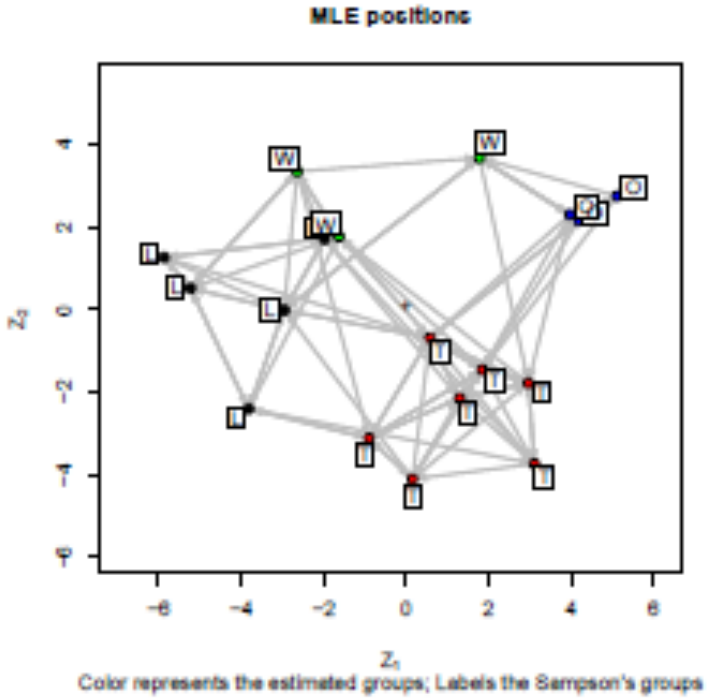
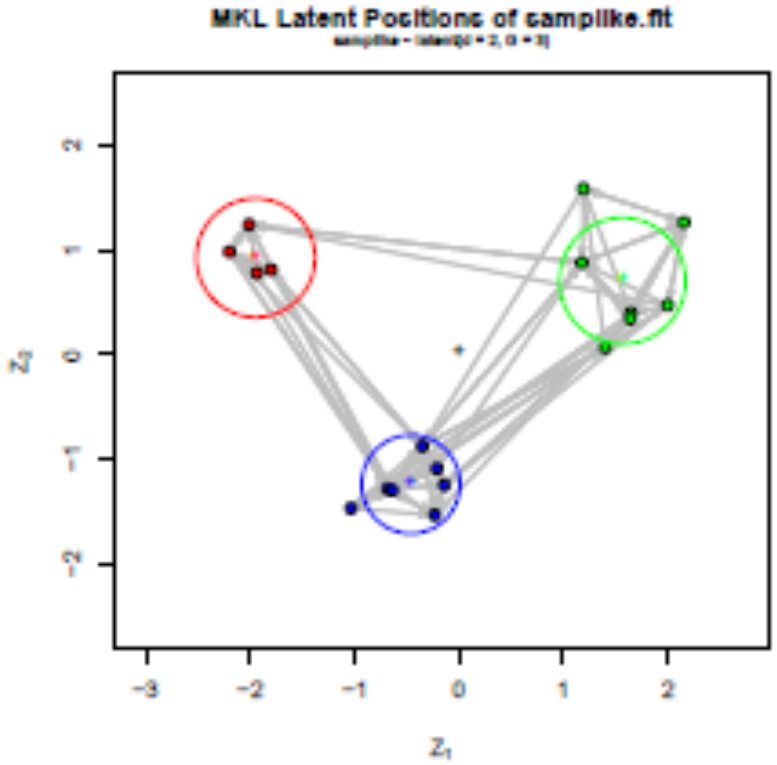
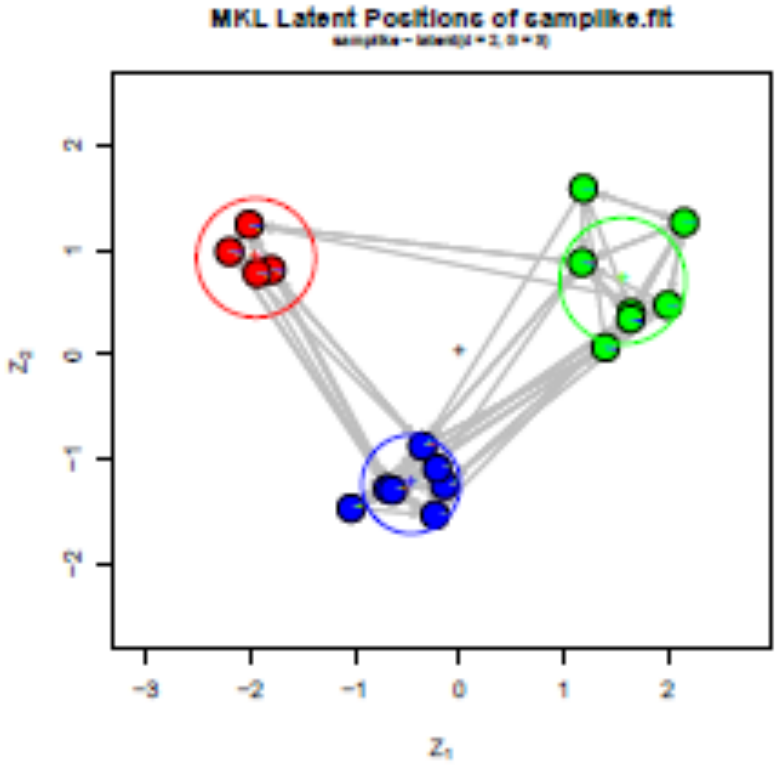


Figure 1: MLE positions for a fit on Sampson's Monks

```
> samplike.fit <- ergmm(samplike ~ latent(d = 2), tofit = c("mle"))
```



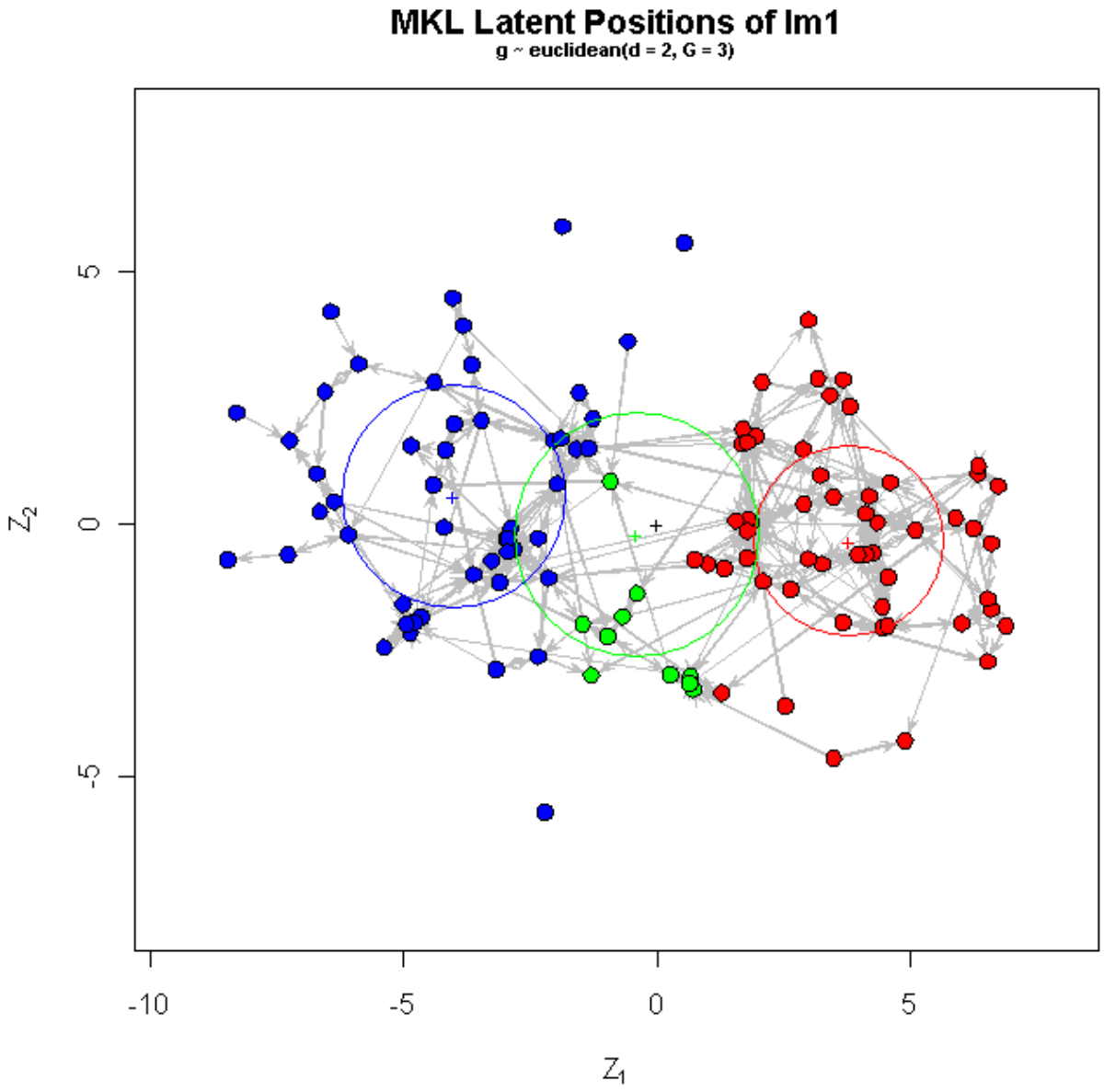



```
> samplike.fit <- ergmm(samplike ~ latent(d = 2, G = 3),
+     verbose = TRUE)

> plot(samplike.fit, pie = TRUE, vertex.cex = 2.5)
```

Example with the Prosper data, with three groups

Dimension isn't just Euclidean. There's some clustering soaking up variation.

Here we assume 3 groups.



Latent space models tend to be (a) much more robust to model specification errors than are ERGMs and (b) have better known convergance properties (i.e. you can prove that the models will converge, which follows because you're making a conditional independence assumption that's not made in ERGM).

But, you rarely know what the dimensions mean socially. *So it provides a fit, but doesn't test a mechanism.*

This is a key difference; if you're goal is out of sample prediction or simply controlling the "noise" of a network, a latent space model is probably the best solution. If your goal is to test a particular network mechanism, an ERGM is probably better.

AMEN: Additive & multiplicative effects from latent factor models (Hoff & Volfovsky)

Basic social relations model

$$y_{ij} = \beta_d^t x_{d,ij} + \beta_r^t x_{r,i} + \beta_c^t x_{c,j} + a_i + b_j + \epsilon_{ij}$$

Dyad effects | Row effects | Column effects | Row error | Col error | dyad error

More general frame:

$$y_{ij} = \beta_d^t x_{d,ij} + \beta_r^t x_{r,i} + \beta_c^t x_{c,j} + a_i + b_j + u_i^t v_j + \epsilon_{ij}$$

Latent multiplicative covariance

Model is very general; can deal with y on any scale (binary to real values), fits latent space & observed covariates.

Computationally intensive…for networks > 100;

Package 'amen'

May 25, 2017

Title Additive and Multiplicative Effects Models for Networks and Relational Data

Version 1.3

Author Peter Hoff, Bailey Fosdick, Alex Volfovsky, Yanjun He

Description Analysis of dyadic network and relational data using additive and multiplicative effects (AME) models. The basic model includes regression terms, the covariance structure of the social relations model (Warner, Kenny and Stoto (1979) <DOI:10.1037/0022-3514.37.10.1742>, Wong (1982) <DOI:10.2307/2287296>), and multiplicative factor models (Hoff(2009) <DOI:10.1007/s10588-008-9040-4>). Four different link functions accommodate different relational data structures, including binary/network data (bin), normal relational data (nrm), ordinal relational data (ord) and data from fixed-rank nomination schemes (frn). Several of these link functions are discussed in Hoff, Fosdick, Volfovsky and Stovel (2013) <DOI:10.1017/nws.2013.17>. Development of this software was supported in part by NIH grant R01HD067509.

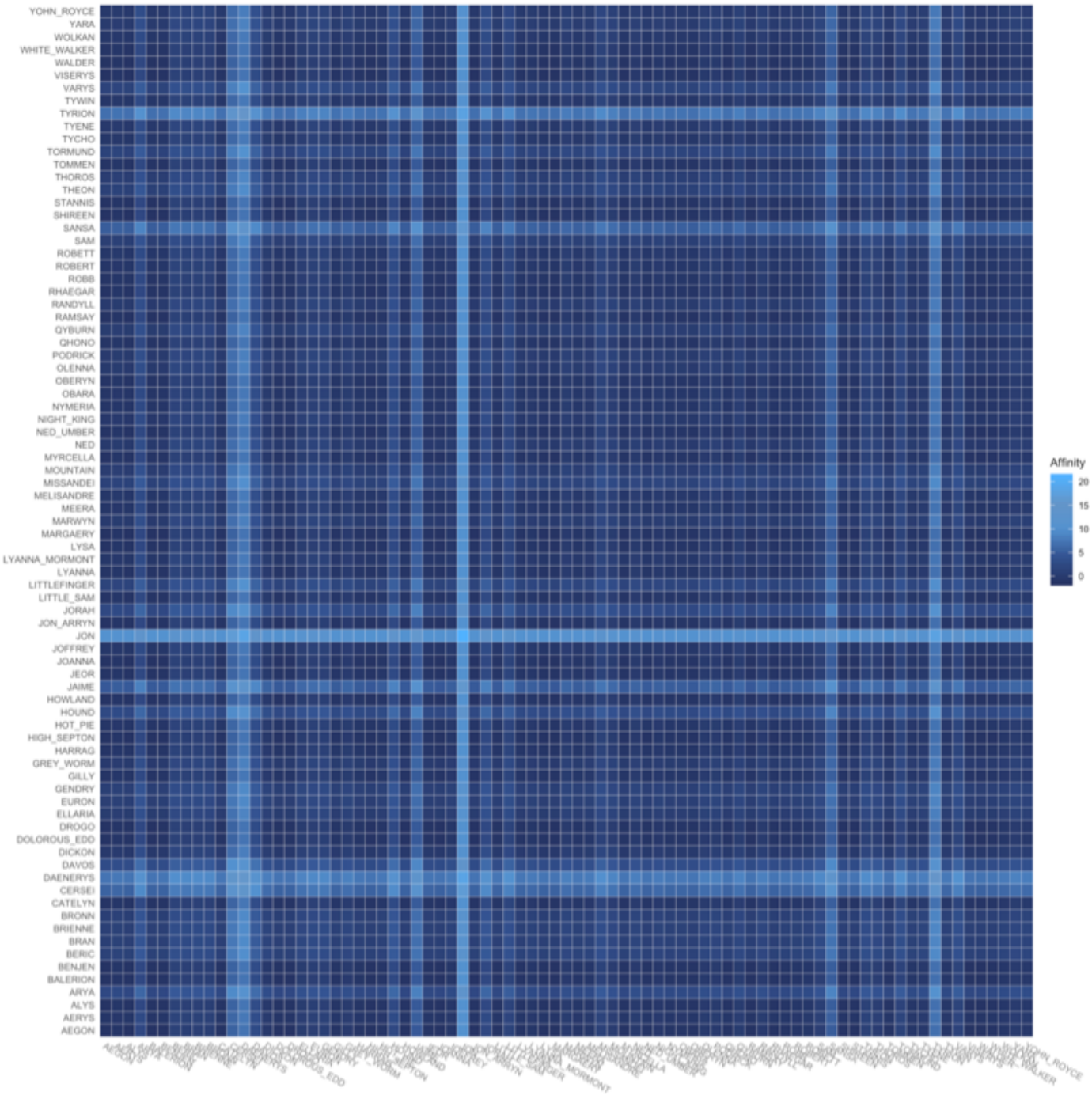
Maintainer Peter Hoff <peter.hoff@duke.edu>

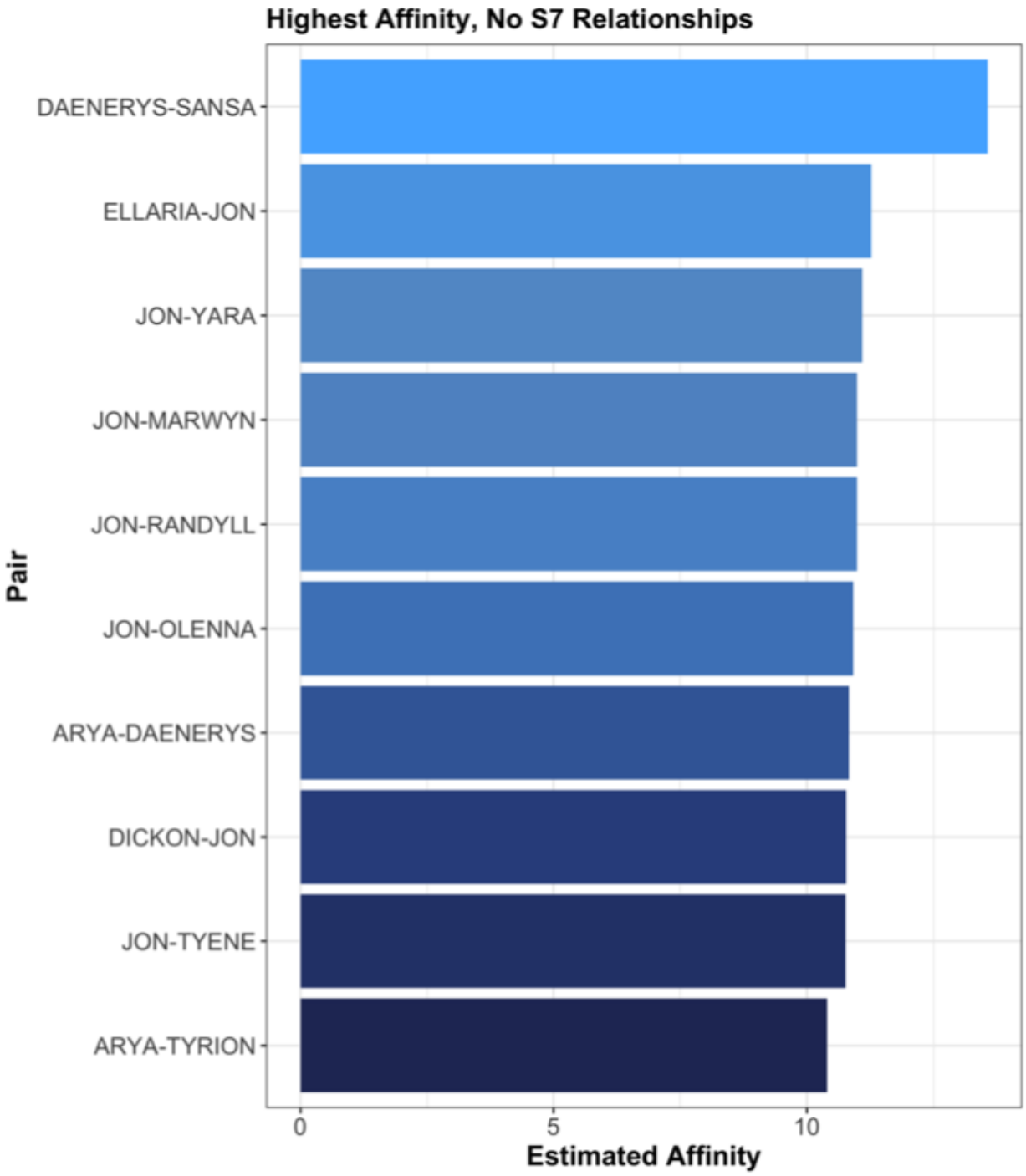
License GPL-3

Date 2017-05-23

LazyData true

Affinity of GoT characters




Highest Affinity, No S7 Relationships
DAENERYS-SANSA
ELLARIA-JON
JON-YARA
JON-MARWYN
JON-RANDYLL
JON-OLENNA
ARYA-DAENERYS
DICKON-JON
JON-TYENE
ARYA-TYRION
Pair
0
5
10
Estimated Affinity

(Field) Experiments

Randomizing into conditions, done by experimenter or naturally by exogenous shock.

Three examples

1. **Peer Effects**: does j influence the behaviour or outcomes of i?

2. **Network Formation**: what conditions whether j forms a tie with i?

3. **Designing networks**: which network structures maximize network level outcomes?