University of Essex

Universidad Católica del Uruguay
Departamento de Ciencias Sociales y Políticas

# Social Network Analysis

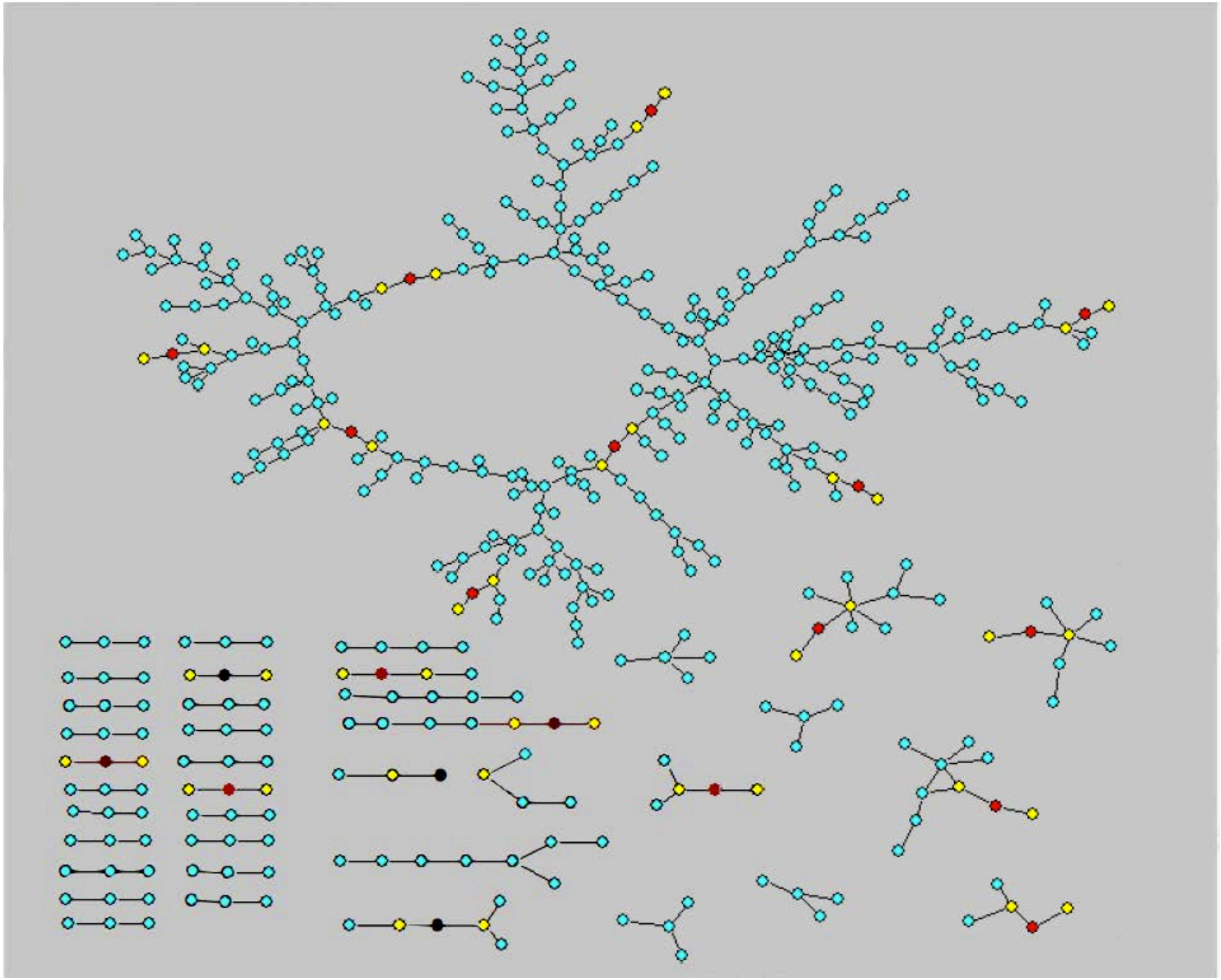
## Day 5 - Data Collection, Communities & Hypothesis Testing

ESSEX SUMMER SCHOOL in Social Science Data Analysis

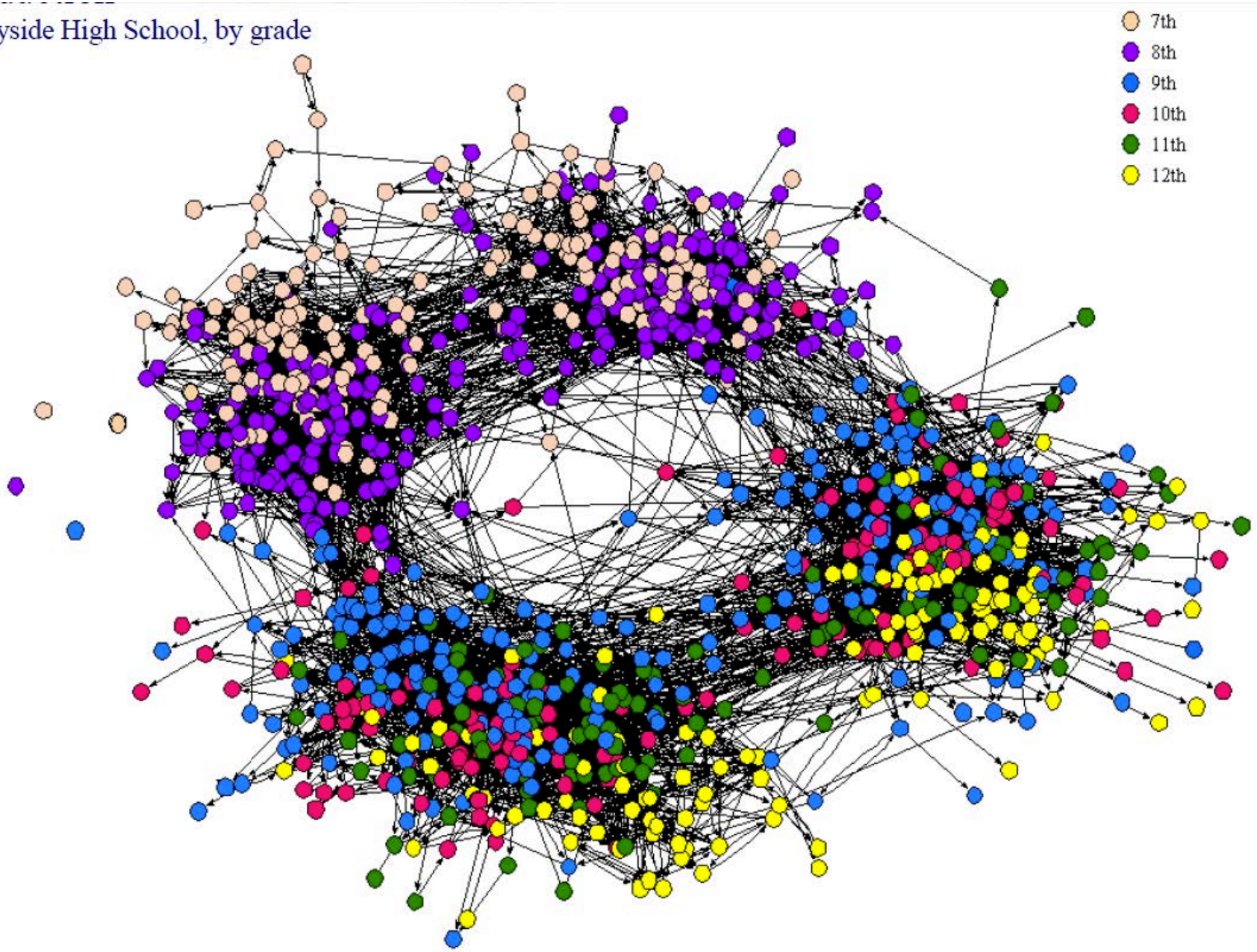
# Part I - Data Collection

"For the last thirty years, empirical social research has been dominated by the sample survey. But as usually practiced, …, the survey is a sociological meat grinder, tearing the individual from his social context and guaranteeing that nobody in the study interacts with anyone else in it."

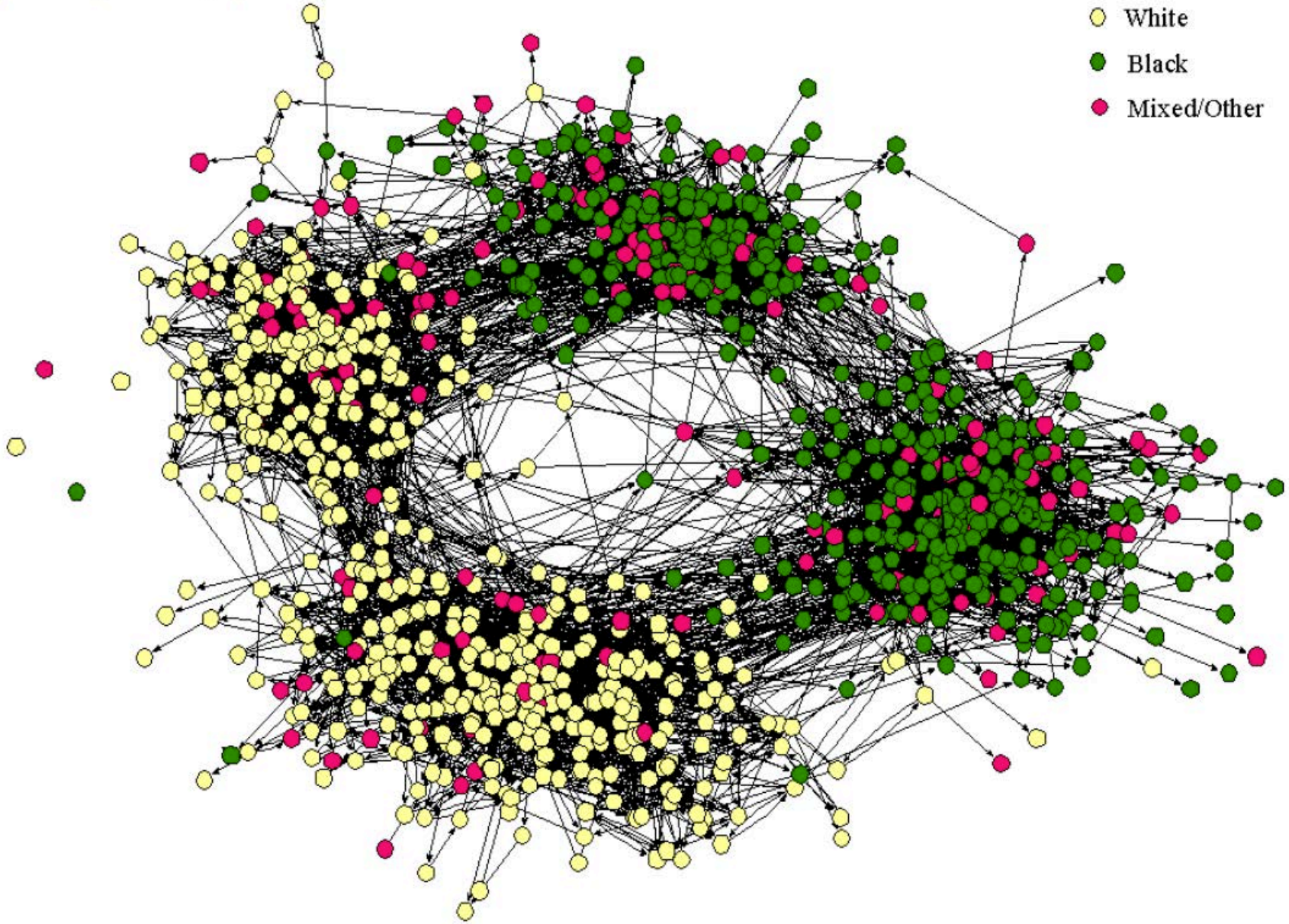
*Allen Barton, 1968 (Quoted in Freeman 2004)*



Countryside High School, by grade



Countryside High School, by race



# Structure Matters

- The structure is real!
  - A more accurate rendering of social reality
- Our job is to try to detect structure and represent it through abstractions
  - Visual representations
  - Mathematical summaries
- Thus, validity is the key research goal

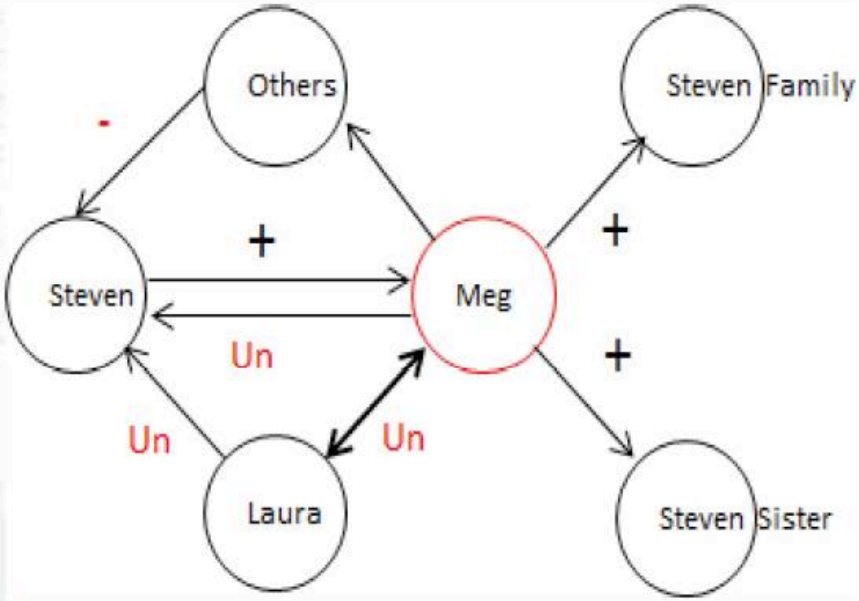
# Structure Matters

- SNA Core Research Goals
  - (1) Accurately represent social structures (descriptive)
    - Implications for outcomes (i.e. health)
  - (2) Explain how social structures come about, and what their consequences are (explanatory)
    - Ties forming and unforming
    - Actual measured outcomes (flows, productivity, good things/bad things)

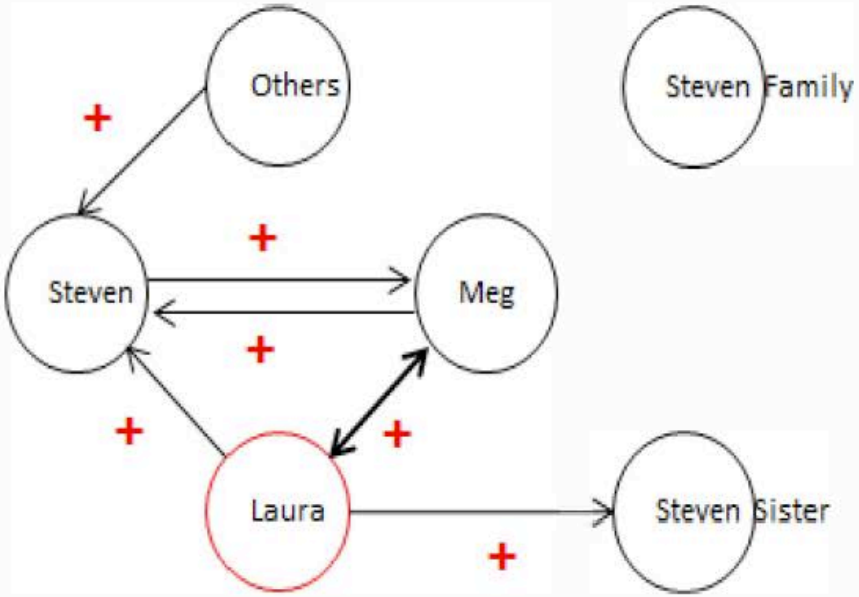
- Network data is everywhere because social structure is everywhere!

| | | |
|---|---|---|
| 1 | Meg | … Do you know Steven Johanson? Alot of people think he's a geek, I |
| 2 | | guess. But he likes me and he's so nice. We talk on the phone alot and I |
| 3 | | went over to his house last night. Nothin' happened but he is really nice |
| 4 | | and his family is nice, and he has a huge house and a pool. (Asshole! J/K) |
| 5 | | His sister is pretty, she doesn't look 12 ½. She looks like she should be in |
| 6 | | 9th grade. A lot of people told me not to worry about what other people |
| 7 | | think. I asked him to TWIRP ["The Woman Is Required to Pay"-Dance] |
| 8 | | (kind of). I still have to figure out what's happening. I don't know what |
| 9 | | we'd do or where we'd go or who with. You're probably thinking I'm |
| 10 | | crazy to go out with Steven, I hope you don't think he's a big nerd cuz I |
| 11 | | know he's not super popular or anything, but not alot of people really |
| 12 | | know him, and once you get to know him, he's super nice. Anyway, |
| 13 | | better go. W/B very soon. |
| 14 | Laura | I know Steven pretty well, he's a great guy. I think it would be awesome |
| 15 | | if you 2 went to TWIRP. He is just shy, not a big nerd, Sarah [his sister] is |
| 16 | | really pretty, we play tennis together. |

## Meg's View




## Laura's View



# Data Collection is Already Theory

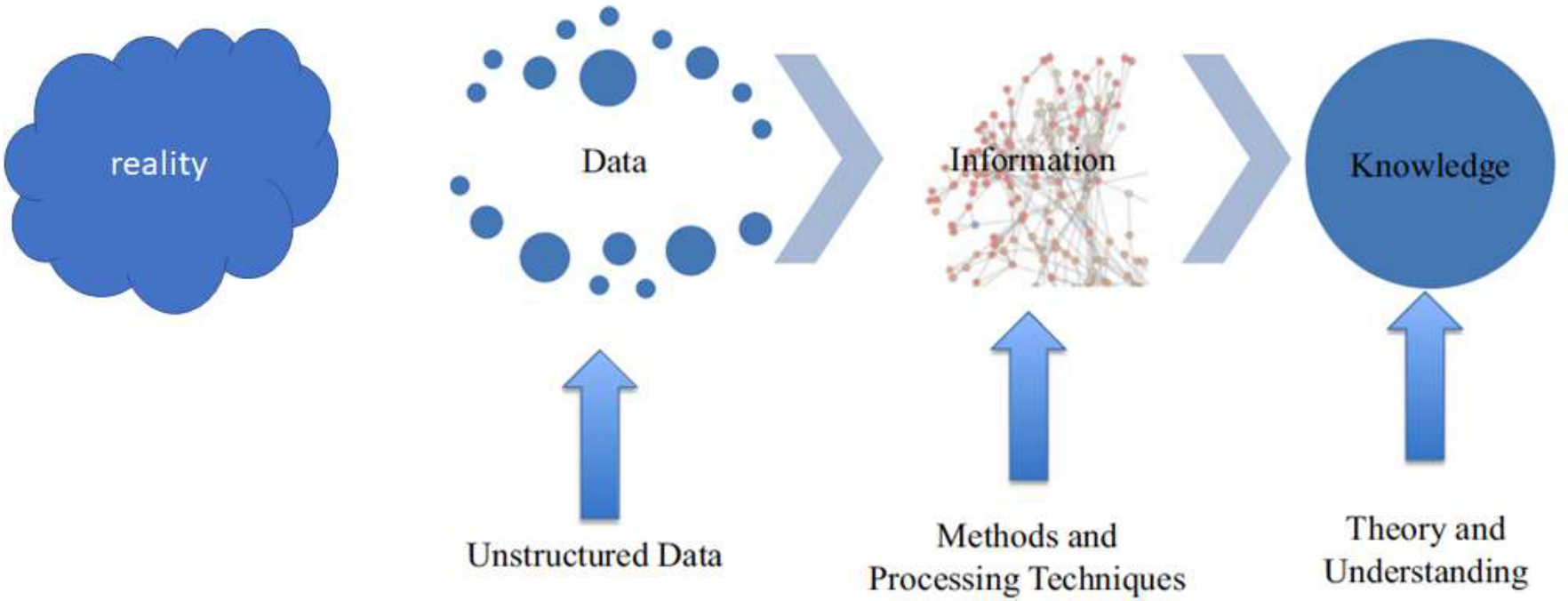



**Fig. 1** Schematic of Data, Information, and Knowledge

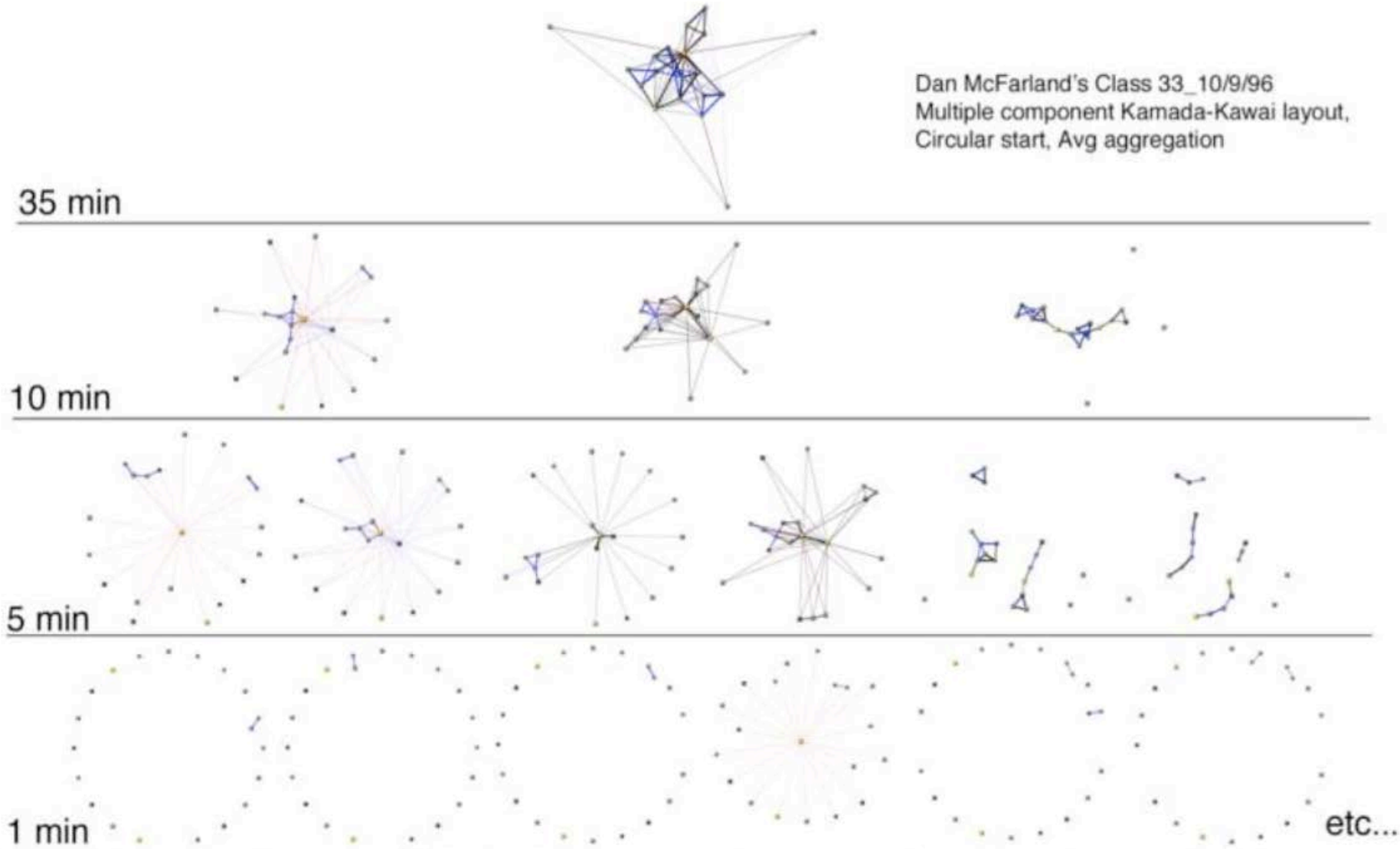



*Figure 1. Interaction data from McFarland's classroom observations viewed at various levels of time aggregation from 35 minutes (one entire class period) to 1 minute (two to three turns of interaction).*

# How to detect structure

- Data Sources

- Most common
  - small group questionnaires,
  - large-scale surveys,
- Less common
  - face-to-face observations,
  - sensor data
- Trendy
  - “scraping” many thousands of websites,
  - using API’s and digital archives.

# How to detect structure

- **Archival Data** – increasingly common!
  - **Easy and cheap data:** easy to scrape, growing in prevalence, longitudinal…
  - **BUT Lots of issues swept under rug…**
    - *Tie construct validity* - What is a tie? Is it really the same type of tie?
      - » Example: coauthoring = are collaborations of N=2, 3, 500 same sort of tie
      - » Example: citations can be used for many reasons (e.g., homage to pioneers, disputing prior work, identifying methods, giving veneer of legitimacy, etc
    - *Identity disambiguation* issues - What is a node?
      - » Who is whom when many have identical names? How do we trace names changes…
    - Websites *contextualize activity* (like a survey or task) and transactional traces reflect *variable participation*. (double ugh)
      - » Can you compare persons who spend 1 min on site to those who many hours? ~Sampling each 1 vs 10000 times.

# How to detect structure

## Observation data

- **Audiovisual**
  - Location in room (field of vision and hearing)
  - Hard to assess who addresses whom
  - Noise
  - Strength - reanalysis
- **Sensor/Wifi**
  - Technical challenges
  - Proximity and exposure is accurate
- **Hand recording** via short hand (McFarland 1999; Diehl and McFarland 2012, Gibson 2001)
  - Accuracy and bias issues of reporter
  - Location in room (field of vision and hearing)
  - Codes specific to theory

- There is no single right way to collect network data! It is always a matter of data availability, strategic tradeoffs, and suitability to your specific theoretical and substantive interests.

- In other words, it's social research.

# Roethlisberger and Dickson 1939

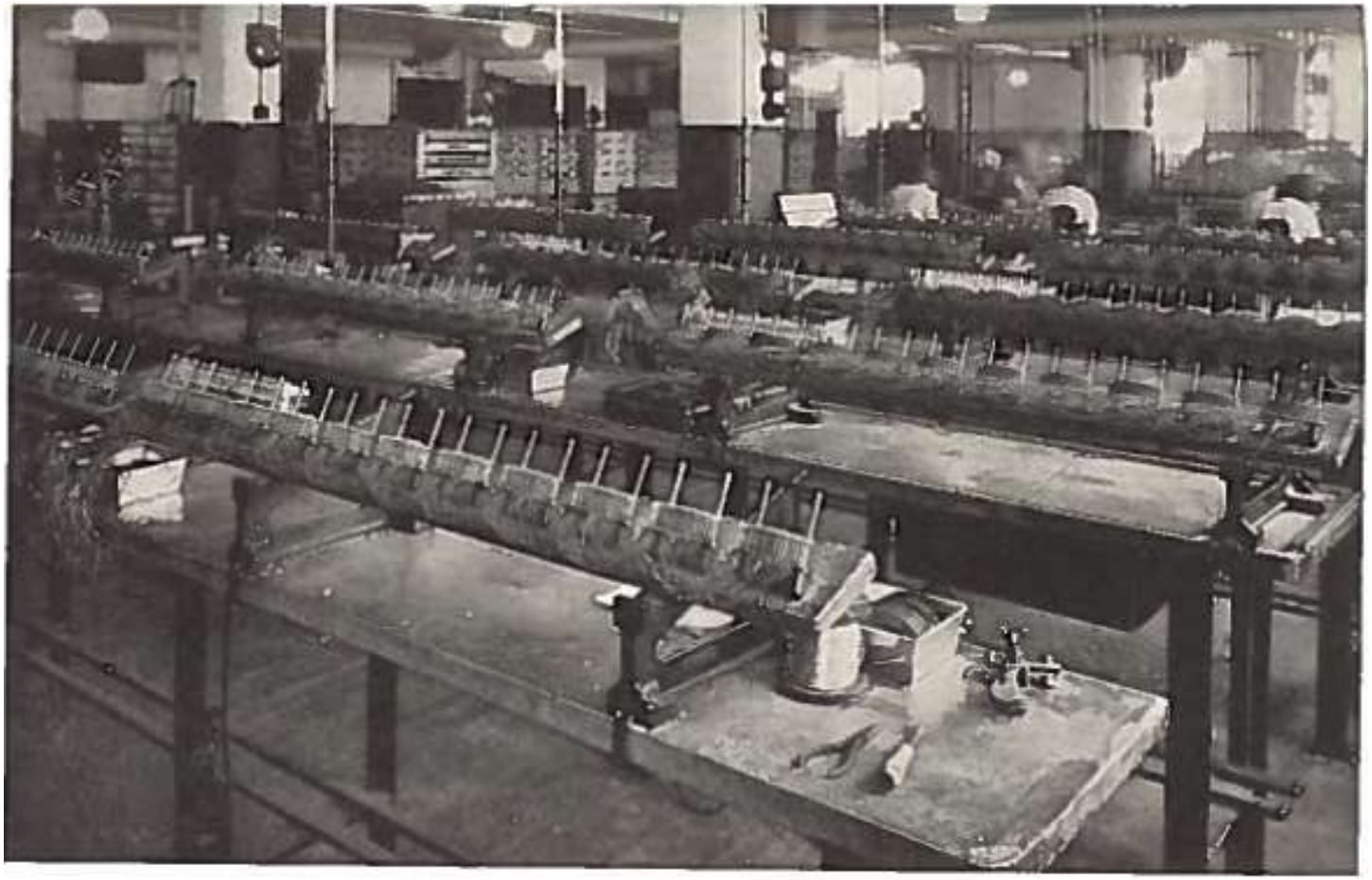


FIGURE 33
PHOTOGRAPH OF A SECTION OF THE BANK WIRING DEPARTMENT, SHOWING BANKS AT DIFFERENT STAGES OF COMPLETION

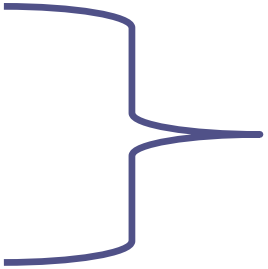
- Clearly, a single room in a plant is not a complete network, as these individuals likely had many friendships outside that room, even at the same plant. However, because the outcome of interest for the research team concerned work productivity, the flows of interpersonal influences that were most likely to bear on this outcome were those in the immediate work environment.

# Types of Network Questions
# Shape Data Collection

| | Networks As Cause | Networks As Result |
|---|---|---|
| Connectionist:<br><br>*Networks as pipes* | Diffusion<br>Peer influence<br>Social Capital<br>“small worlds” | Social integration<br>Peer selection<br>Homophily<br>Network robustness |
| Positional:<br><br>*Networks as roles* | Popularity Effects<br>Role Behavior<br>Network Constraint | Group stability<br>Network ecology<br>“Structuration” |

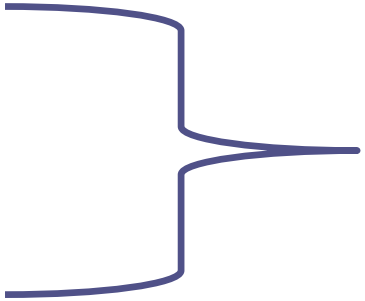
# How Do Networks Form?

- Key Processes
  - Homophily
  - Shared Foci



Exogenous Factors

  - Reciprocity
  - Transitive Closure
  - Preferential Attachment



Endogenous Factors

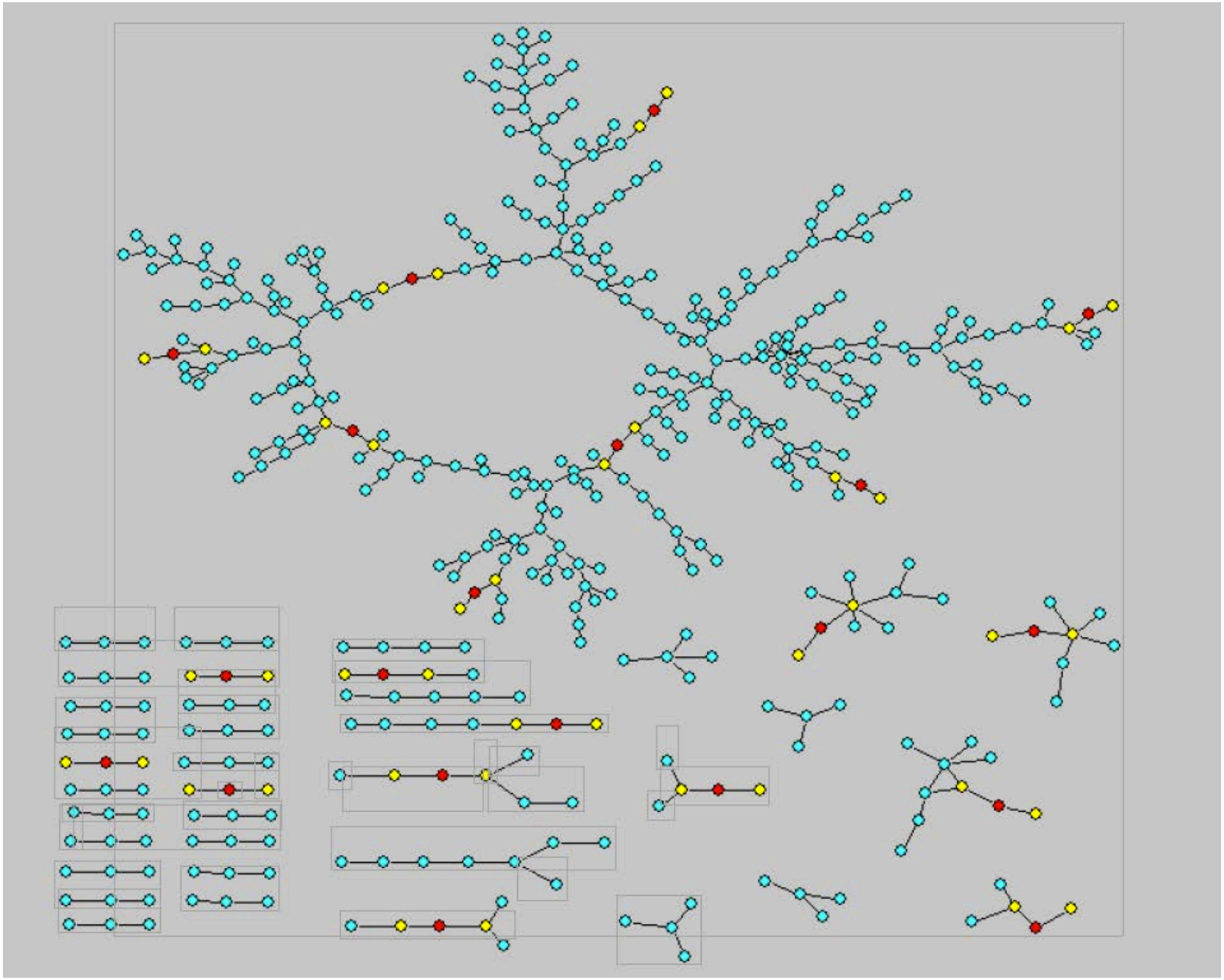
# Defining Nodes & Ties

- **Kinds of actors** (nodes, vertices, points)
  - People, groups, organizations, communities, nations
    - Often include information on demographics, behaviors, and attitudes of actors.

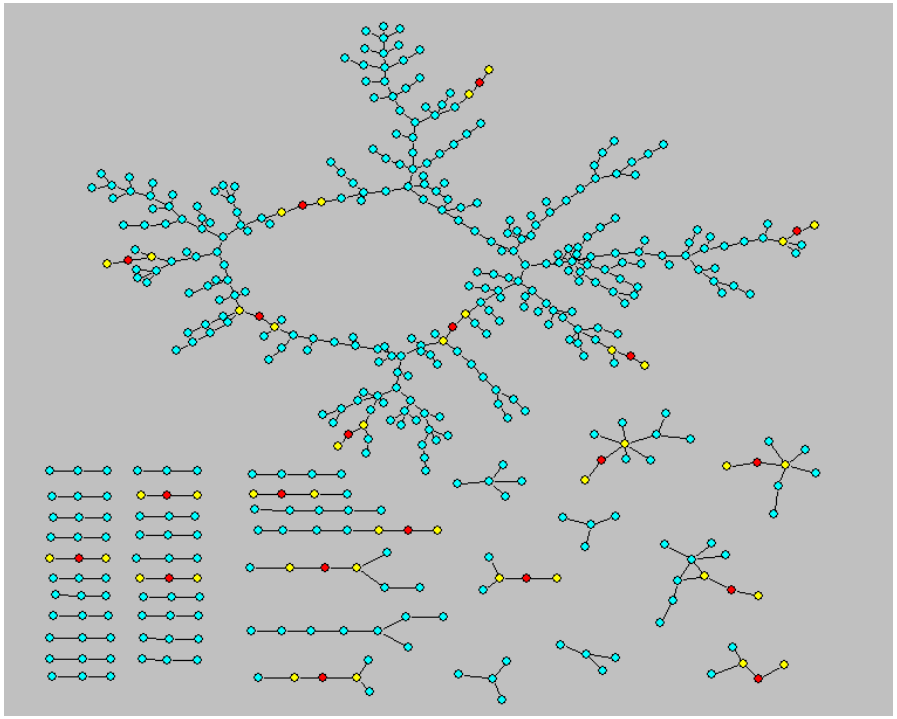
- **Levels of Analysis**
  - Individual ego, dyad, triad, clique/group/role, whole social structure

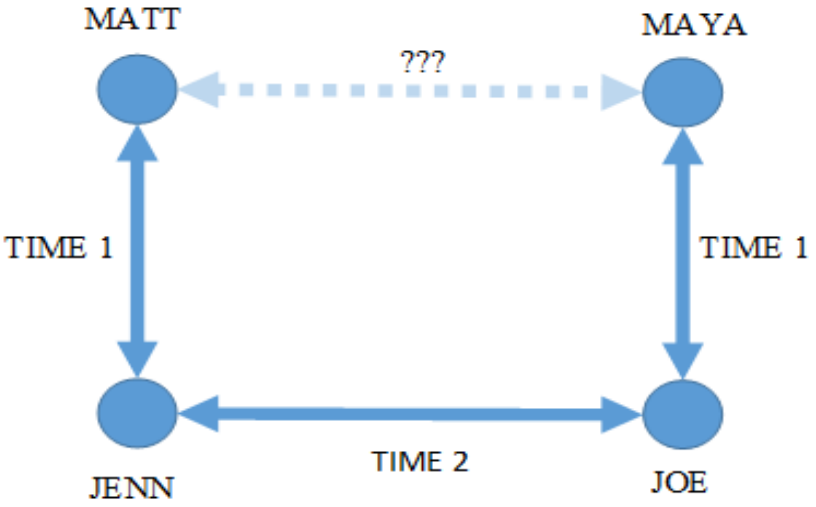
- **Units of time**
  - Seconds, minutes, hours, days, weeks, months, years, decades, centuries

What dyadic/triadic processes generated this network?



# Inductively Uncovering “Rules” of Interaction






Romantic “Leftovers”: dating the ex of your ex's current partner.

# What ties do you want to collect data on?

- **Similarities** in which nodes are located in the same regions in physical and social space (same neighborhoods, same department, same club).
- **Relations** in which nodes operate within a system of roles (e.g., father of; friend of; teacher of, etc.) and have cognitive or affective orientations toward one another (likes, dislikes, admires, etc.).
- **Interactions** in which concrete interactions occur between nodes (advice, romance, bullying, etc.).
- **Flows** in which nodes transfer some material or cultural object, goods, information, or influence (ideas, beliefs, practices, etc.)

# Network Qualities

- **Forms of data:**
  - Relational network 1-mode (sociometric) – who to whom (e.g., friends)
  - Affiliation networks 2-mode (memberships) – who to what (e.g., club affiliations).
  - Cognitive networks – all relationships seen from each participant

# Questions

- **Consider your interests and the sort of data you have or would like to have:**
  - **What sort of network questions interest you? Connections or roles?**
  - **What sort of data do you think you need to answer these questions?**
    - **Local or Complete?**
    - **Directed or Undirected?**
    - **Cross-sectional or longitudinal?**
    - **One-mode or two-mode?**

# Data Collection Instruments

# Survey and Questionnaire Design

(Marsden 1990, 2005)

- **Name Generator Surveys**
  - **Free choice** (as many as you like) vs **Fixed choice** ("only top five")
    - Free >> Fixed choice: Issue of artificial cap – limited to 5 friends
    - Order reported is interesting
  - **Roster** (full list of classroom or school) vs **Recall** (up to respondent)
    - Choice has recall issues – memory / cold-call listing not always complete so you may get false negatives.
    - Rosters are preferred method as it relies on recognition instead of recall – but it may induce false positives.

## Local / Ego Network Data

When using a survey, common to acquire “ego-networks” or local network information. Three parts to collection:

- 1. Elicit list of names - “Name Generator”
- 2. Get information about each person named
- 3. Ask about relations among persons named

# Social Network Data

*Sources - Survey*

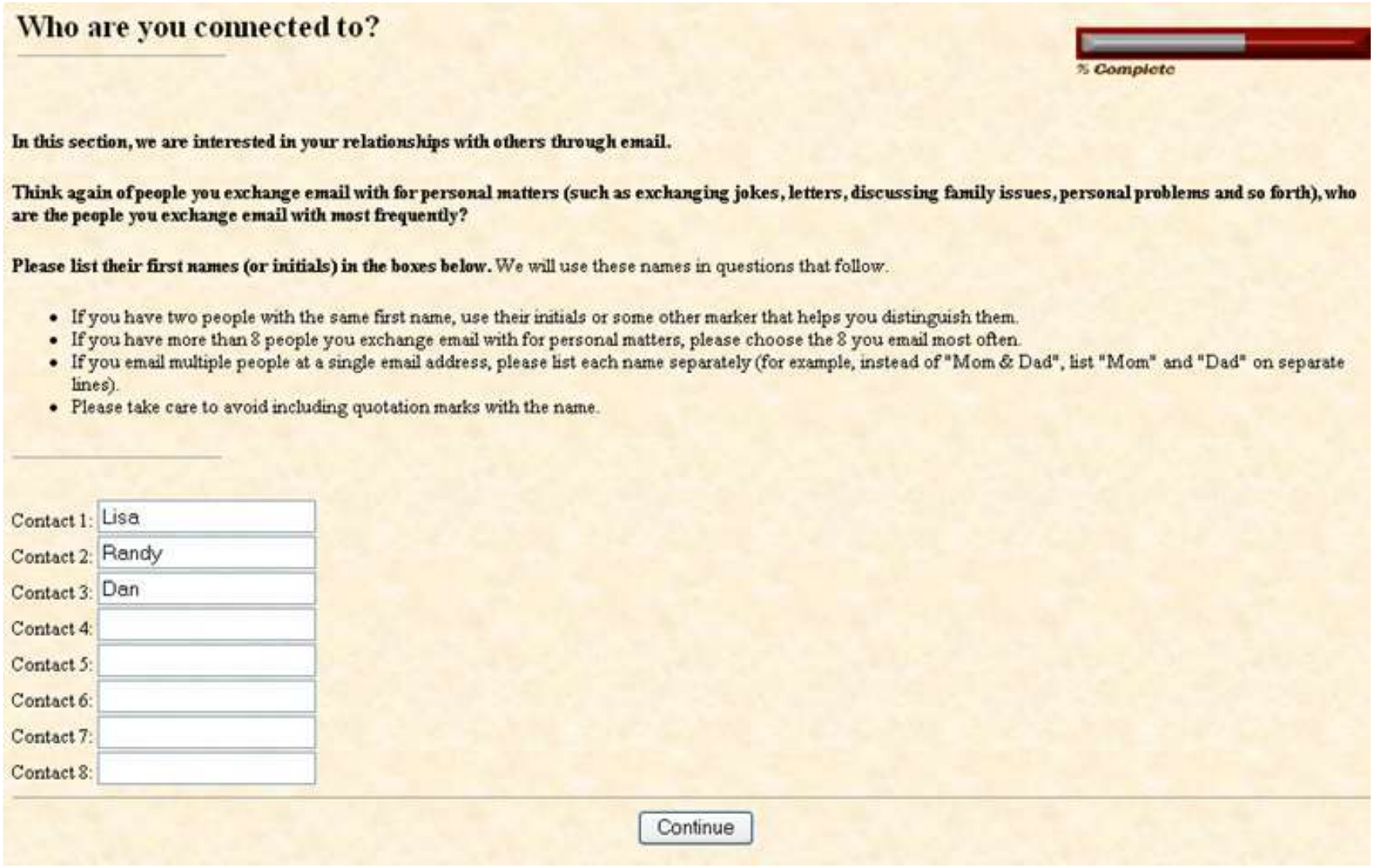
a) Network data collection can be time consuming. It is better (I think) to have *breadth* over *depth.* Having detailed information on <50% of the sample will make it very difficult to draw conclusions about the general network structure.

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

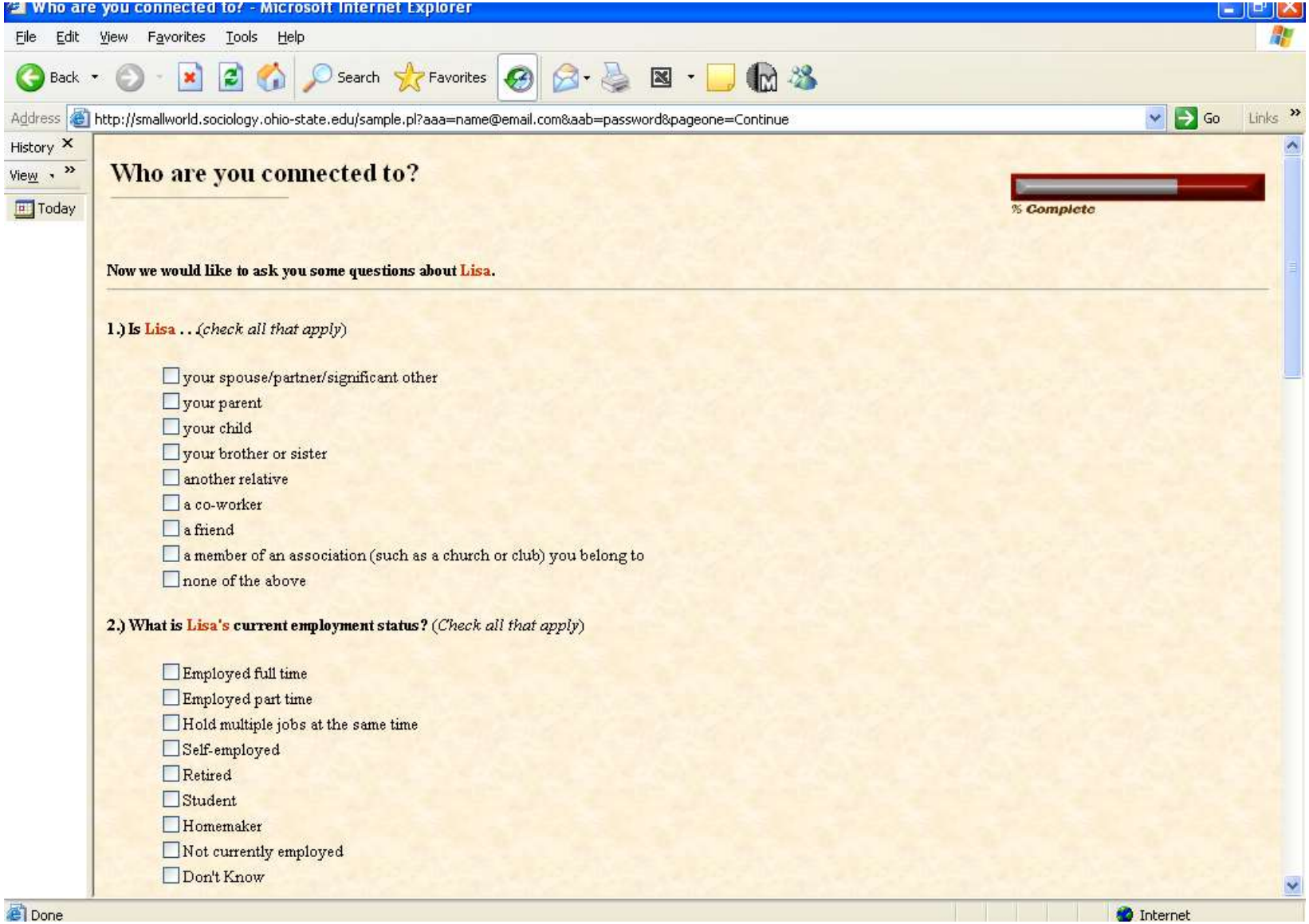
c) It is common to limit people to a small number if nominations (~5). *This will bias network measures*, but is sometimes the best choice to avoid fatigue.

d) People answer the question you ask, so be clear in what you ask.

Part 1
Electronic Small World name generator:

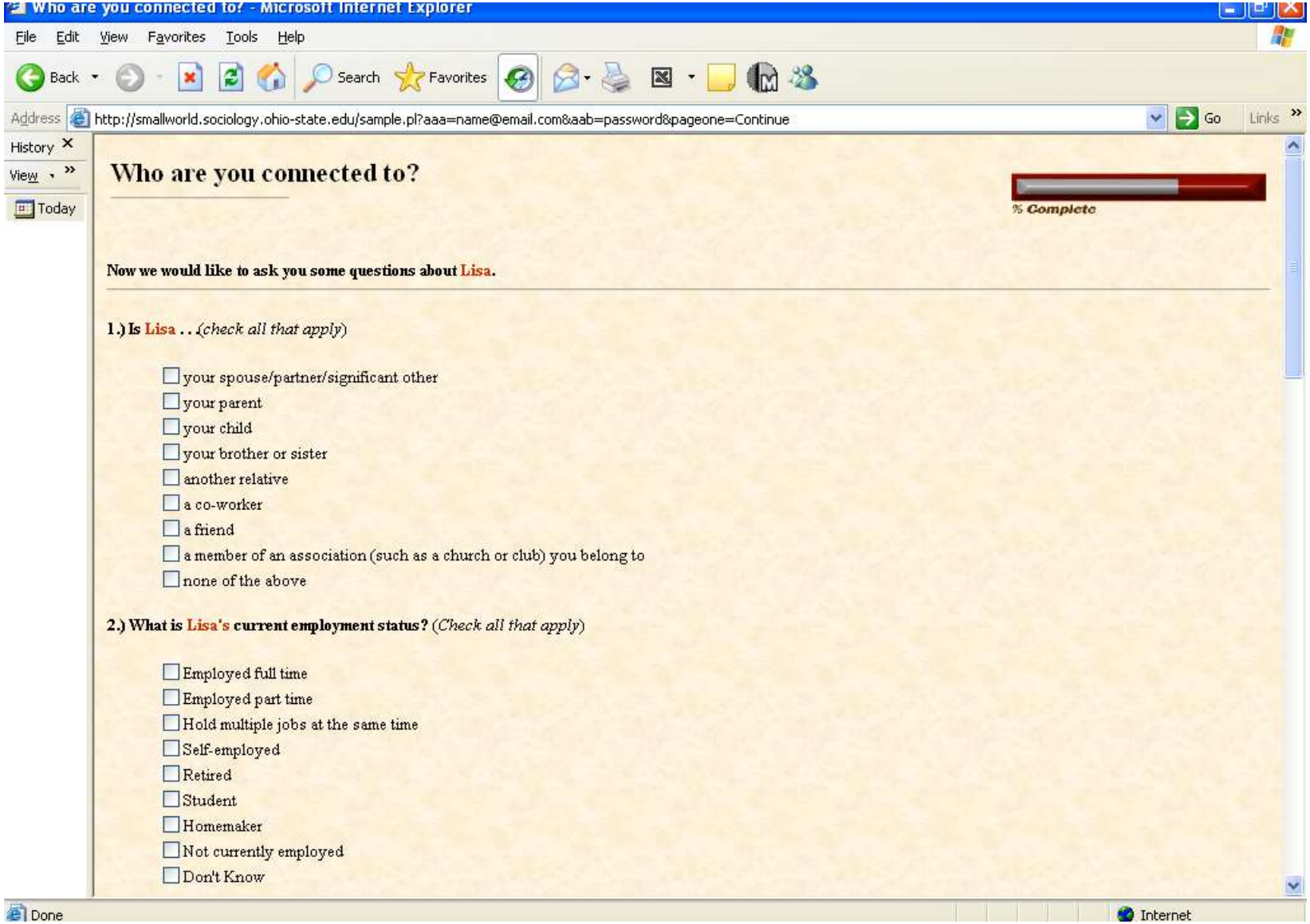


The second part usually asks a series of questions about each person




Will generate N x (number of attributes) questions to the survey

The second part usually asks a series of questions about each person




Will generate N x (number of attributes) questions to the survey

## Complete Network Data:

To acquire complete network data, you need to collect information on “all” relations within a specified **boundary**.

- Requires sampling every actor in the population of interest (all kids in the class, all nations in the alliance system, etc.)

- Two general formats:
  - Recall surveys (“Name all of your best friends”)
  - Roster formats: Give people a list of names, have them check off those with whom they have relations.

## *Friends Nomination Form* -- Who are your close friends that you usually hang around with? Please list only as many people as you usually hangout with.

| 1. | 2. | 3. | 4. | 5. In what settings do you usually see this friend? *For each friend check as many as apply* | | | | | | | 6. When do you see this friend? *Check as many as apply* | | | 7. | 8. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| What are your friends full names? *Please print their first and last names* | About how old is this friend? | How long have you been friends? | Is this friend male or female? *Check Male or Female* | In My School Classes | In a School Activity (like a team or extra-curricular) | In a Non-School Club or Activity (like a youth group, or church) | At Work | In My Neighborhood | In my family | Other | Less than Once a week | Weekdays | Weekends | Do you know this friend's parents? *Check Yes or No* | Is this friend a best friend? *Check Yes or No* |
| Example: Jane Doe | 16 yr. | 6 mos. | ___ Male<br>_X_ Female | *X* | *X* | *X* | | | | | | *X* | *X* | _X_ Yes<br>___ No | ___ Yes<br>_X_ No |
| (a) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (b) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (c) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (d) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (e) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (f) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (g) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (h) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (i) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |
| (j) | | | ___ Male<br>___ Female | | | | | | | | | | | ___ Yes<br>___ No | ___ Yes<br>___ No |

**(1) Who do you regularly talk to in this class?**
**(2) Who do you ask for help with schoolwork in this class?**
**(3) Who do you joke around and socialize with in this class?**
**(4) Who do you consider a friend you can share personal stuff with?**
**(Check all the boxes that apply)**

| *Name of Classroom Member* | *(1) Regularly talk to* | *(2) Go to for help with work* | *(3) Joke around & socialize with* | *Consider a friend* |
|---|---|---|---|---|
| EXAMPLE: Joe Bloe | X | | X | |
| Teacher | | | | |
| Student 1 | | | | |
| Student 2 | | | | |
| Student 3 | | | | |
| Student 4 | | | | |
| Student 5 | | | | |
| Student 6 | | | | |
| Student 7 | | | | |
| Student 8 | | | | |
| Student 9 | | | | |
| Student 10 | | | | |
| Student 11<br>Student 12 | | | | |
| Student 13 | | | | |
| Student 14<br>Student 15 | | | | |
| Student 16 | | | | |
| Student 17 | | | | |
| Student 18 | | | | |
| Student 19 | | | | |
| Student 20 | | | | |
| Student 21 | | | | |
| Student 22 | | | | |
| Student 23 | | | | |
| Student 24 | | | | |
| Student 25 | | | | |

# Key issues

- Whole network designs need good response rate – say, 90%
- We want truthful data
- As a result …
  - Careful attention to questionnaire design
    - Length, question wording, attractiveness
  - Work to build trust
  - Work to inspire interest
  - If you want to collect network data from the same location ever again, handle the data ethically and carefully

# What to ask about

- Depends entirely on the research question
- You get to study any kind of tie you want
  - Nose-licking in cows
- At the same time … for any two people
  - You want to know something of the nature of their relationship
    - Which can be multiplex
  - Something of the amount of interaction they have

# Roster vs Write-in

**Roster method (closed-ended)**

- Boundaries are known and all actors listed
- Becomes cumbersome as networks grow in size
- Fewer concerns about respondent recall and accuracy
- Each actor has approximately an equal chance of being selected

**Write-in method (open-ended)**

- More subject to recall error
- Can use a fixed choice method limiting the number of actors elicited
- Each actor in the network does not have an equal chance of being chosen given recall and freelisting issues
- Can make getting valued ties more complicated
- Better for face-to-face interviews where probing can be used

# Serial vs parallel

- Serial (repeated)
  - Focuses attention on the tie
  - Tends to keep definition of "friend" the same across all alters
- Parallel (grid)
  - May focus respondent's attention on the alter as a whole
  - More halo effects, less control over tie definitions

## Repeated Roster

Q1. Please indicate which of the following you would converse with if you met them on the street.

Demi Moore ☐
Jennifer Anniston ☐
Michael Douglas ☐
David Bowie ☐
Bob Dylan ☐

…..

Q2. Please indicate which of the following people with whom you work.

Demi Moore ☐
Jennifer Anniston ☐
Michael Douglas ☐
David Bowie ☐
Bob Dylan ☐

….

## MultiGrid

Q1 Using the checkboxes below, please indicate those people you **would converse with if you met them on the street**.

Q2. Check off the names of the people you **work with.**

Q3. Check off the names of a selected set of people whom you don't know but **would like to know**, based on things you heard, or their interests, etc.

| Name | Q1: Would converse if met on the street | Q2: Work with | Q3: Would like to Know |
|---|---|---|---|
| Demi Moore | ☐ | ☐ | ☐ |
| Jennifer Anniston | ☐ | ☐ | ☐ |
| Michael Douglas | ☐ | ☐ | ☐ |
| David Bowie | ☐ | ☐ | ☐ |
| Bob Dylan | ☐ | ☐ | ☐ |
| Hugh Jackman | ☐ | ☐ | ☐ |
| Kurt Russell | ☐ | ☐ | ☐ |

# Binary or valued?

What do you need to know?
- Nature of the relation
- Amount of interaction

- For relational event type data, you probably need valued data
  - How often you interact with that person
  - Number of emails sent to them
- Properties of a relation
  - You know who is friends with whom, now you want to know how long they've known each other
- For relational states, binary data might be sufficient
  - Who are you friends with?
  - Is this person a co-worker?
- For degree to which an alter satisfies a condition, must make a trade-off
  - To what extent you regard this person as a friend?

# Binary or valued?

**Binary**

- Cognitively easy
  - Fast
  - Resp stays focused
- Limited discrimination
- Lets respondents make own decisions about cutoffs
  - Which may be good or bad

**Valued**

- More nuanced results
- Cognitively difficult
  - Tiring
  - Very slow
  - Results may not be meaningful
- Some network procedures can’t handle valued data

# Asking frequencies or amounts

| Absolute rating | Relative ranking | Sequential choices |
|---|---|---|
| "How often do you talk to each person, on average?"<br>1. Once a year or less<br>2. Every few months<br>3. Every few weeks<br>4. Once a week<br>5. Every day | "How often do you speak to each person on the list below?"<br>1. Very infrequently<br>2. Somewhat infrequently<br>3. About average<br>4. Somewhat frequently<br>5. Very frequently | 1. Who do you talk to at least once every few months? (check all that apply)<br>2. Who do you talk to at least once every few weeks?<br>3. Who do you talk to at least once a week?<br>4. Who do you talk to every day? |
| • Need to do pre-testing to determine appropriate time scale<br>• Danger of getting no variance<br>• Assumes a lot from resps | • Requires less of respondents; easier task<br>• Is automatically normalized within respondent<br>  • Removes response set issues<br>  • Makes it hard to compare values across respondents (in different rows of data matrix) | • Same data as absolute rating<br>  • less tiring for respondent<br>  • But questionnaire may look longer<br>• With online surveys, can pipe responses so that respondent only sees names checked off in previous question<br>• final question will have few names to react to |

# 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

# Sampling & Network Boundaries

- **Sampling**

  (Laumann, Marsden and Prensky 1989)

  - **Position-based** approach – ex: employment in an organization
  - **Event-based** approach – ex: regulars at the beach
  - **Relational** approach based on connectedness – at least two forms:
    - Snowball (Granovetter – start with fixed set and see who connected to them, connected to them, etc).
    - Expanding selection format (Doreian and Woodward 1992) – start with fixed set and see who is connected to them more than once, and add them – should show boundary

## Snowball Samples – Relational Approach:

- Effective at providing network context around focal nodes. Works much the same as ego-network modules. Ask at least some of the basic ego-network questions, even if you only plan to sample (some of) the people your respondent names.

    1. Start with a name generator, then demographic / relational questions
    2. Get contact information from the people named
    3. Have a sample strategy (which listed people to follow up with)
        - Random walk design (Klovdahl)
        - Attribute design (make sure to walk within clusters)
        - Strong tie design
        - All names design (big)
    4. Stopping criteria – usually density cutoff (when it diminishes)

- Issue: tends to form network around starting individuals, so their selection is most important (e.g., elite networks).

# *Defining Network Boundaries*

**Where** does your network begin & end? (Laumann et al 1983)
**When** does your network exist? (Moody et al 2005)

- **Realist** Approach
  - Participants define it via their collectively shared subjective awareness of membership
- **Nominalist** Approach
  - Analyst imposes a conceptual framework to serve their analytical purposes

| | **Realist Approach** | **Nominalist Approach** |
|---|---|---|
| **Static** (Where is a network?) | Classroom, School | Teacher and social worker networks |
| **Temporal** (When is a network?) | Class period, semester, school year | Minutes, hours, months, years |

# Social Network Data

*Level of Analysis*

What *scope* of *information* do you want?

•Boundary Specification: key is what constitutes the "edge" of the network

| | Local | Global |
|---|---|---|
| "Realist" (Boundary from actors' Point of view) | Everyone connected to ego in the relevant manner (all friends, all (past?) sex partners) | All relations relevant to social action ("adolescent peers network" or "Ruling Elite" ) |
| 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 ("friends in school" or "votes on the supreme court") |

# Issues with social networks survey data...

## How *Reliable* are SNA data?

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

# Types of Error

- **Reliability**
  - Do you get stable or consistent reports on ties?
- **Accuracy**
  - Does the measure reflect a real relationship? Is it on target?
- **Recall**
  - Are you getting completeness or capturing all ties in the sample?
- **Precision**
  - Does the measure have exactness?

# Survey Accuracy Issues – does measure reflect concept?

- Inaccuracy from *survey item's design*
  - Rosters force recognition that may not exist (false positives)
  - Recall allows respondent to forget ties (false negatives)

- Inaccuracy from *informant*
  - Respondents tend to see self as central (Kumbassar et al 1994)
  - Accuracy of short term recall of observed ties is 50% accurate (Bernard Killworth and Sailer 1981; Freeman et al 1987). More accurate on *long term* associations.
  - More accurate reports of *reciprocal* / *transitive* / *cliqued* relations than asymmetric / intransitive relations (Kumbassar et al 1994; Freeman 1992).
  - *Central actors* are more competent informants (especially with cognitive networks and accurate depictions of the ties others think they hold).

# 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

Easy:

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

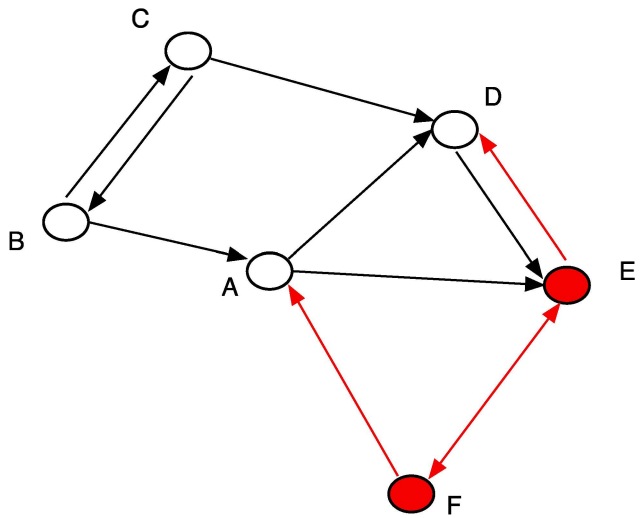
Harder: Impute ties

- 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.

Panel A. True Network with Missing Nodes and Edges Highlighted




Observed Node

Missing Node

Imputed Node

Observed Edge

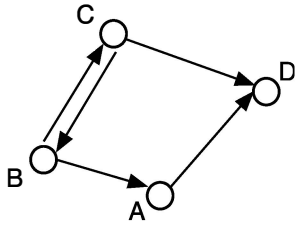
Missing Edge

Imputed Edge

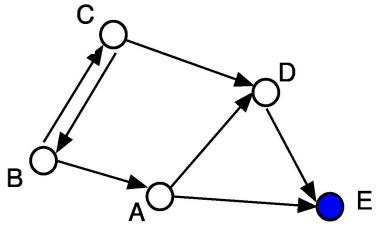
Imputed Edge with probability p, set to observed rate of recipocity (here=.25)

Panel B. Observed Network under Diffrent Imputation Types

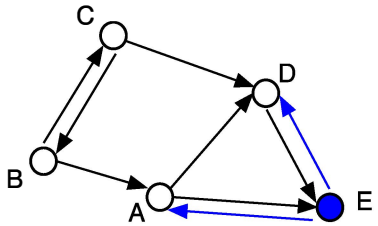
No Imputation (listwise deletion)



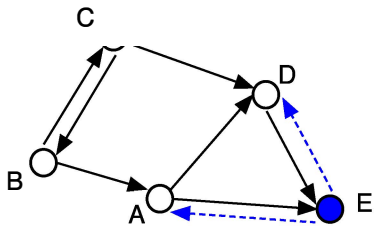

Network Reconstruction with Directed Tie Option




Network Reconstruction with Reciprocated Tie Option




Network Reconstruction with Probabilistic Tie Option



# 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

# Potential Risks Associated with Relational Data

## *Outing People*

Minor: Mom Finds Out Mike Smokes

Major: Wife Finds Out that Her Husband Has Been Cheating

## *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?

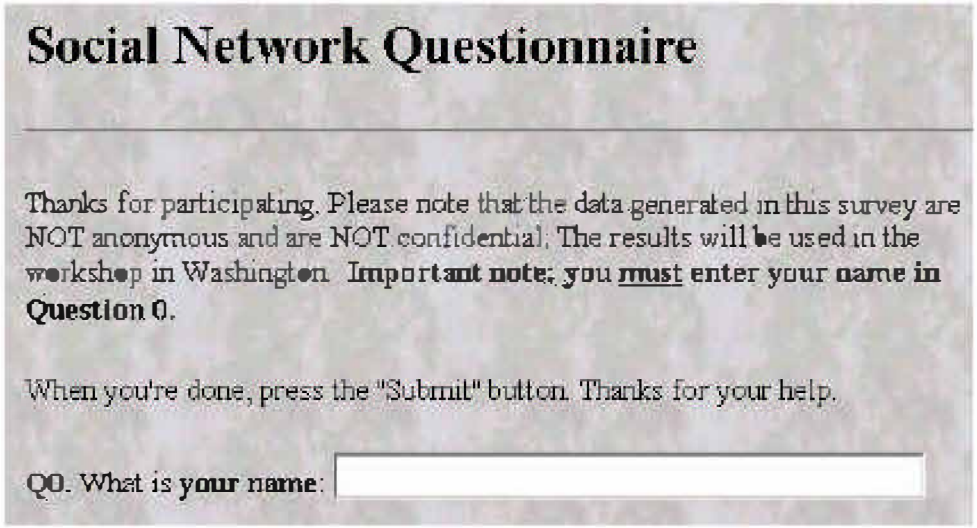
If a known-to-be STD positive person names a partner, do we inform the partner of the respondent's STD status?

## *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**

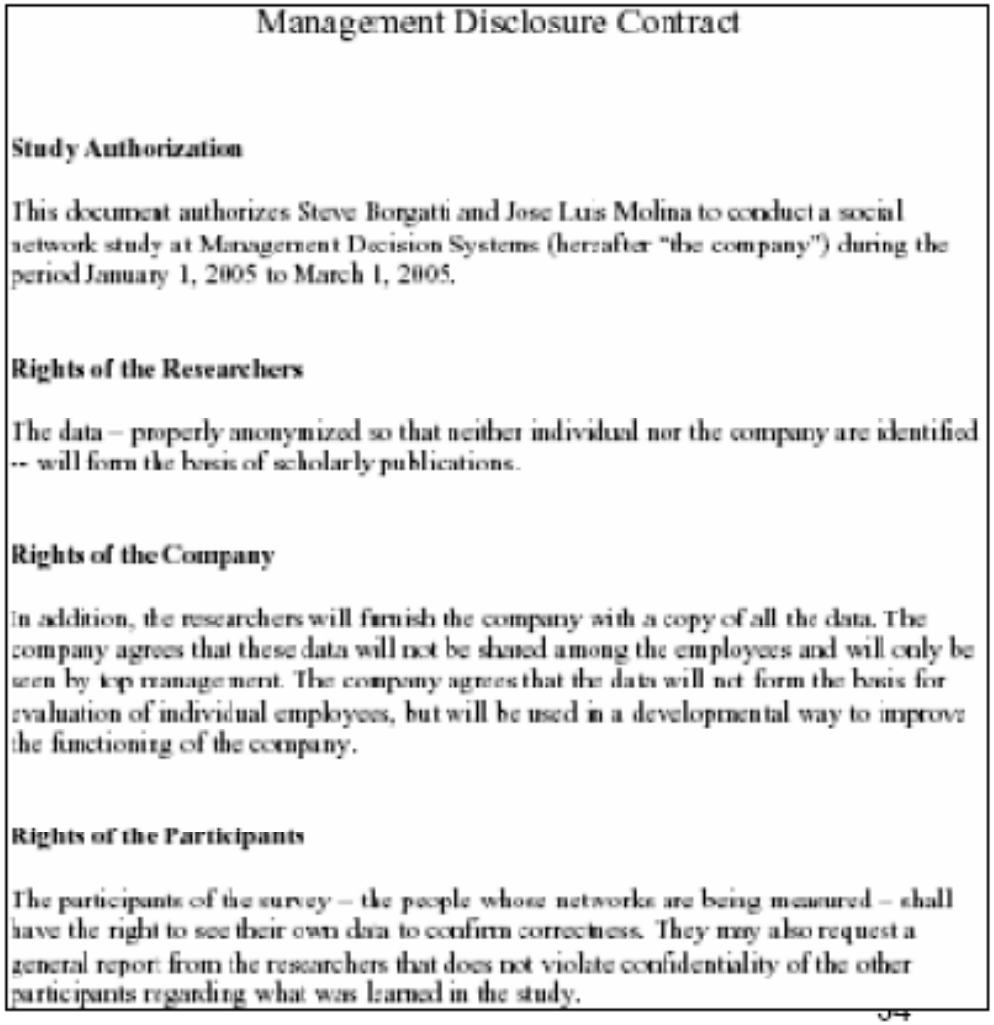
Thanks for participating. Please note that the data generated in this survey are NOT anonymous and are NOT confidential. The results will be used in the workshop in Washington. **Important note: you must enter your name in Question 0.**

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**

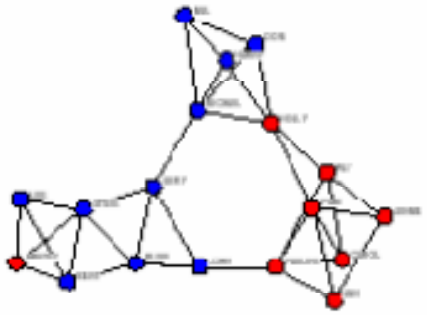
This is a social network study in which we will try to map out the communication network of the organization.

**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

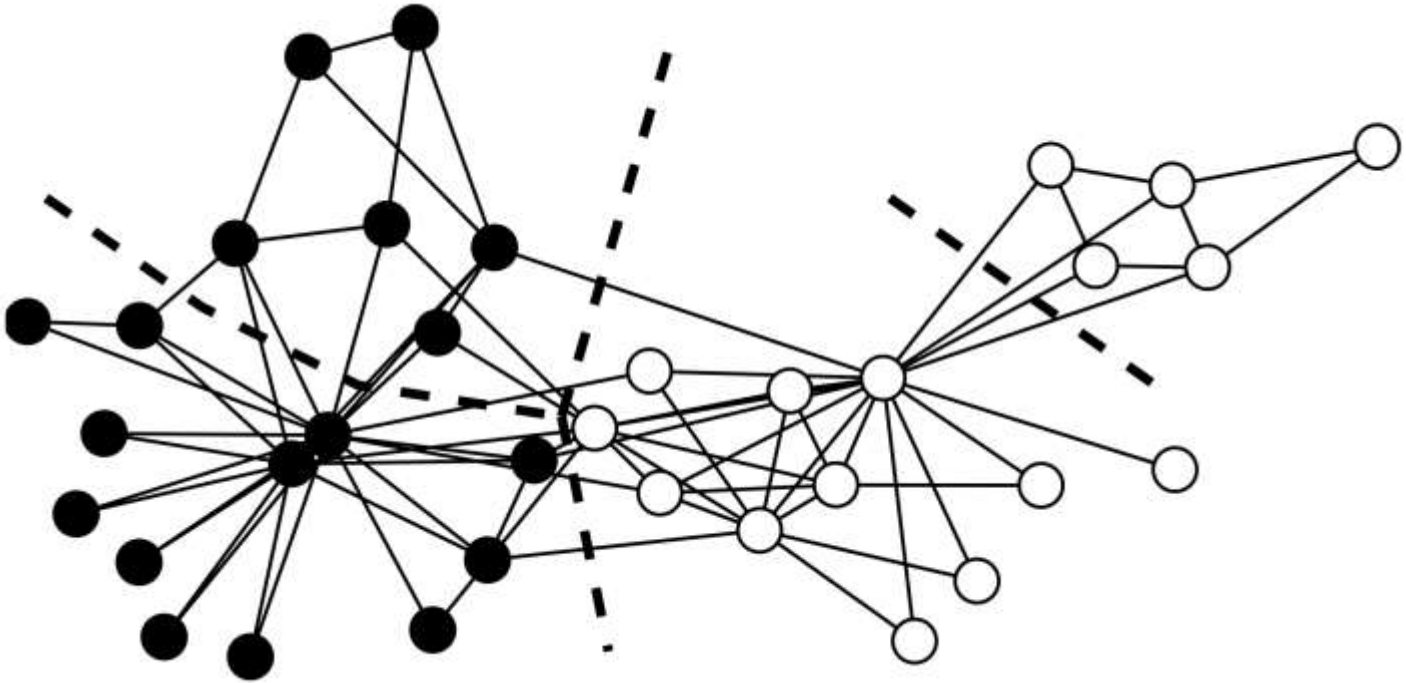
How long will identifying information be retained

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***

# Part II - Subgroups & Communities

# Karate Club Example



This partition optimizes ***modularity***, which measures the number of intra-community ties (relative to a random model)

***"If your method doesn't work on this network, then go home."***

# 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:

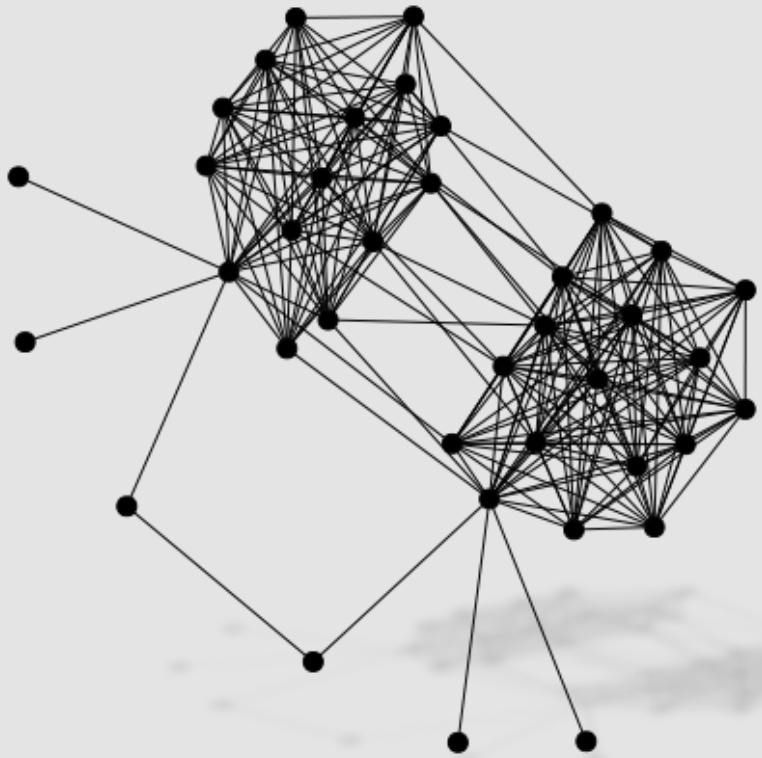
- A community should be densely connected
- A community should be well-separated from the rest of the network
- Members of a community should be more similar among themselves than with the rest

Most common..

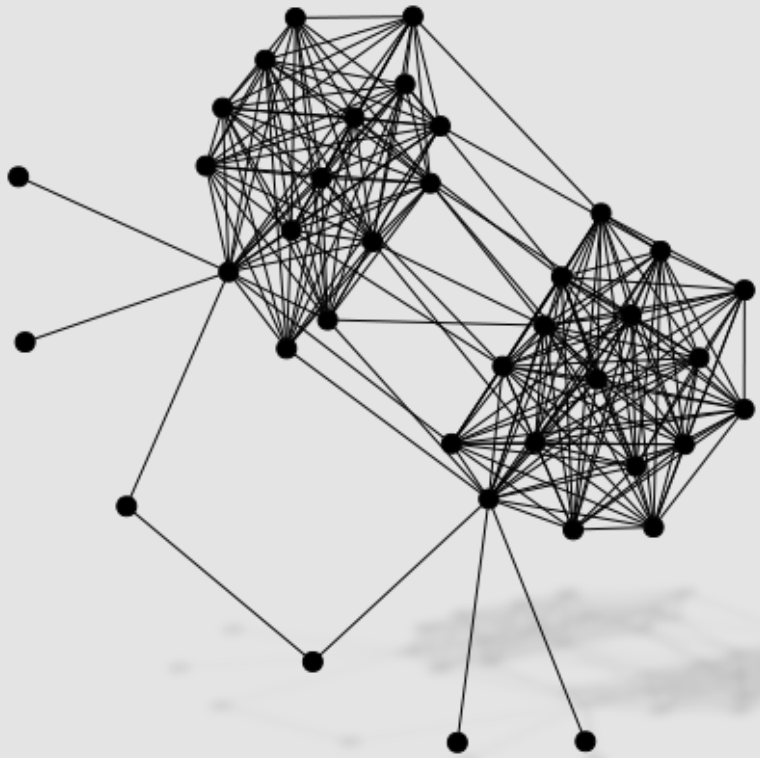
nr. of intra-cluster edges > nr. of inter-cluster edges

**Typology of network communities**

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



**Imagine this Graph ….**

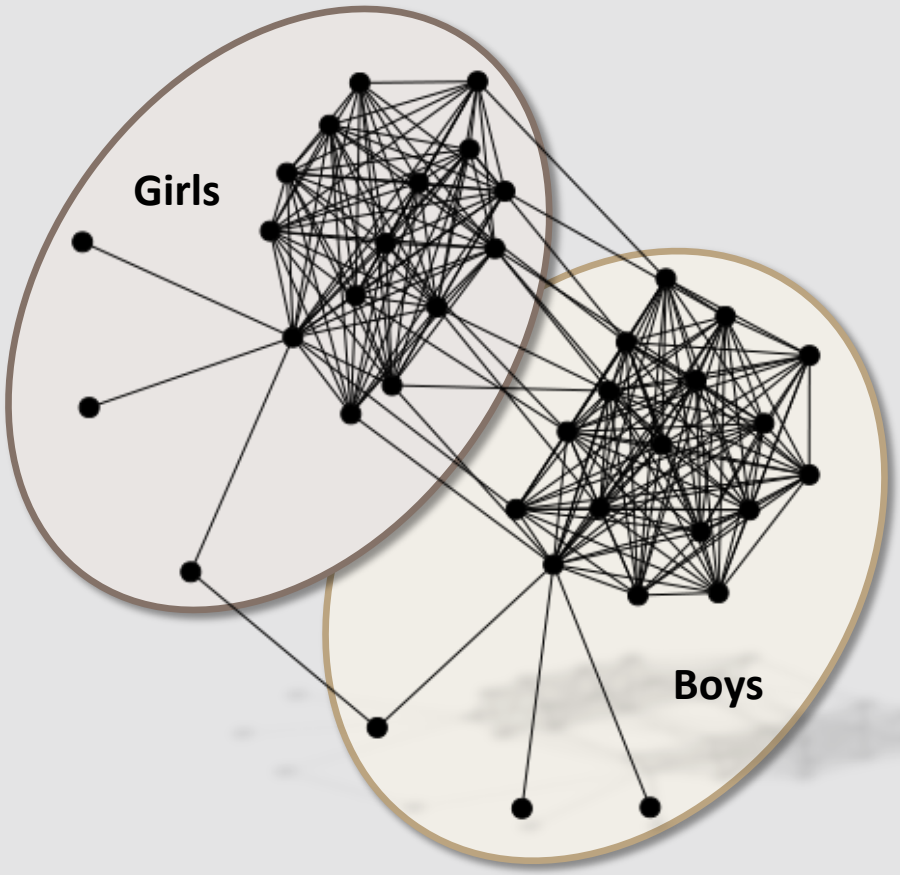


**Vertices**: People
**Edges**: Friendship

**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?**

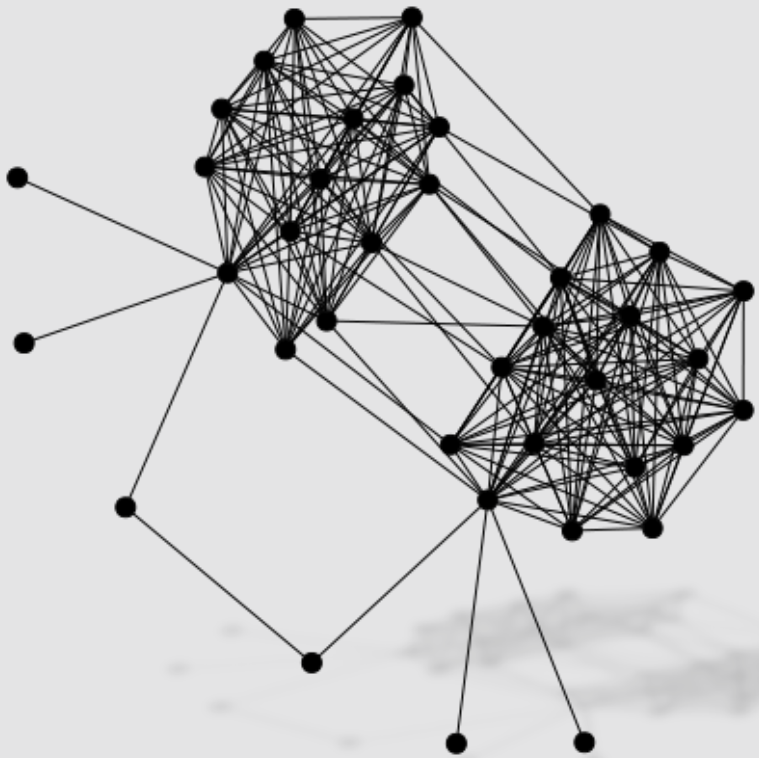



**Vertices**: People
**Edges**: Friendship

**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?**

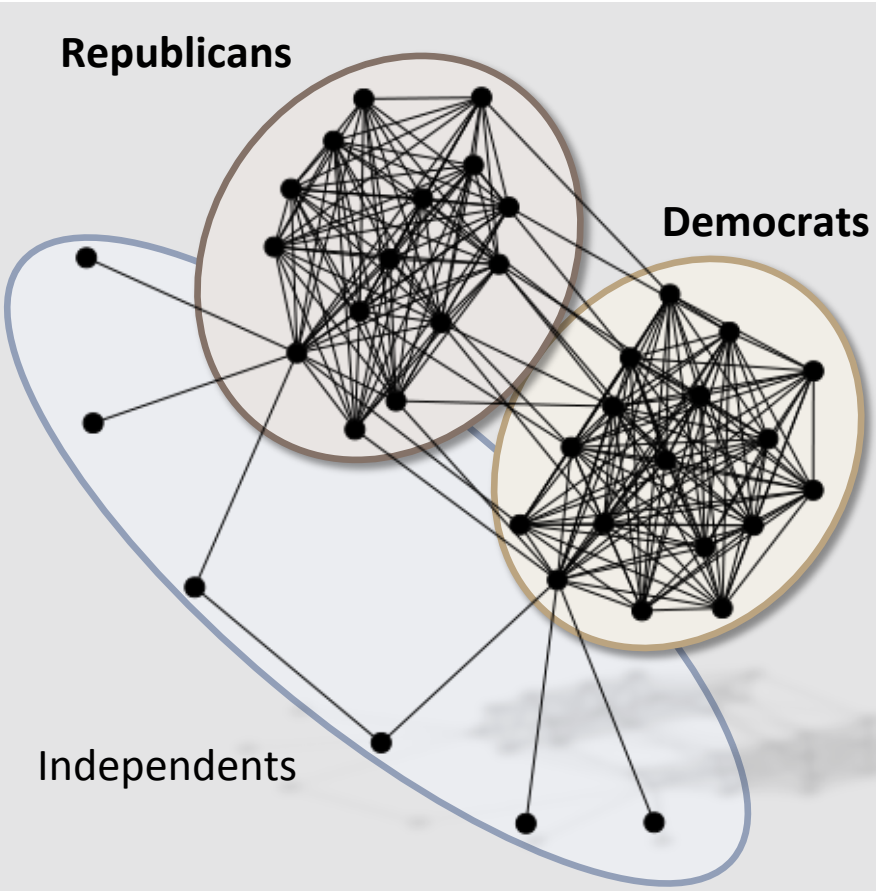


**Vertices**: People
**Edges**: Co-voted
at least once

**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.**

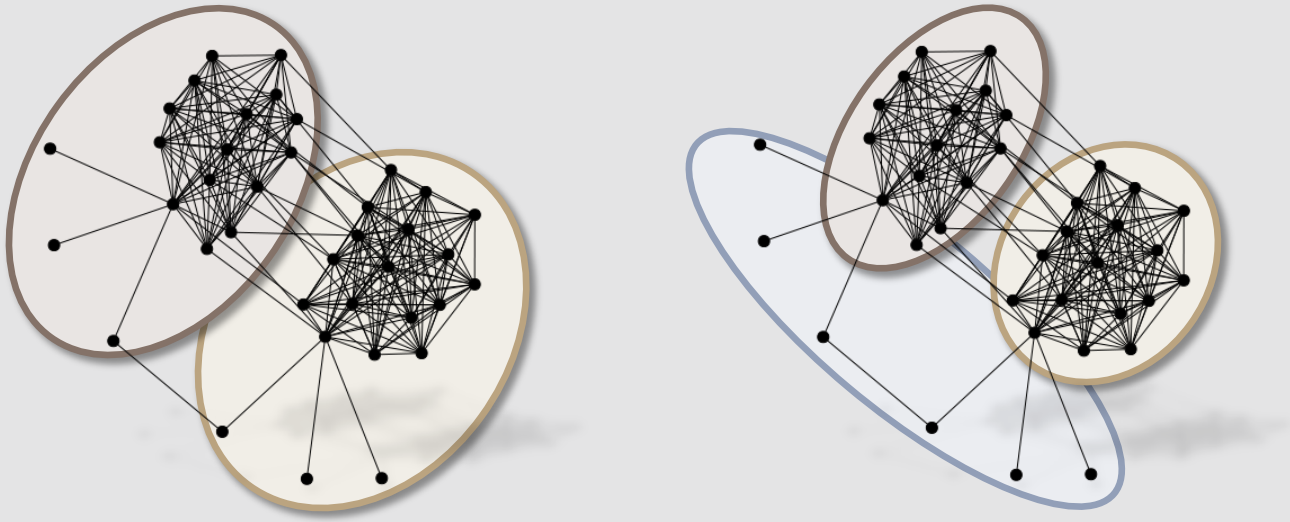
**Ideas:** Issue position, geography, ethnicity, gender

**How might we assign communities to this network?**

# context matters

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

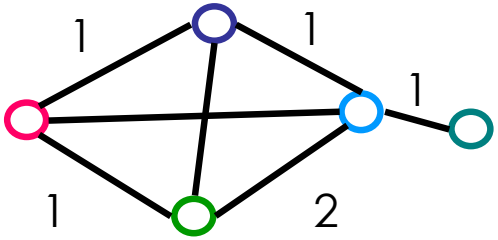
**Community detection is not a concept that can be divorced from context.**

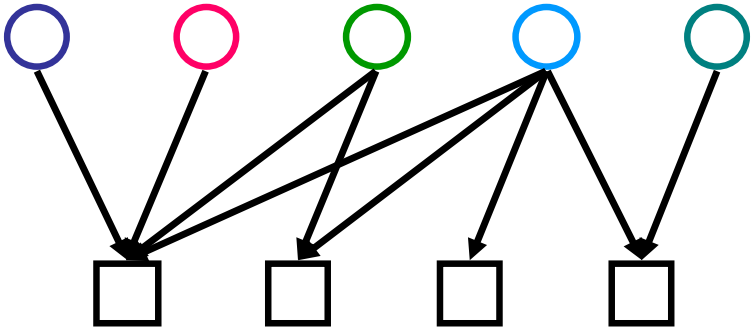


# context matters – why do we observe communities at all?

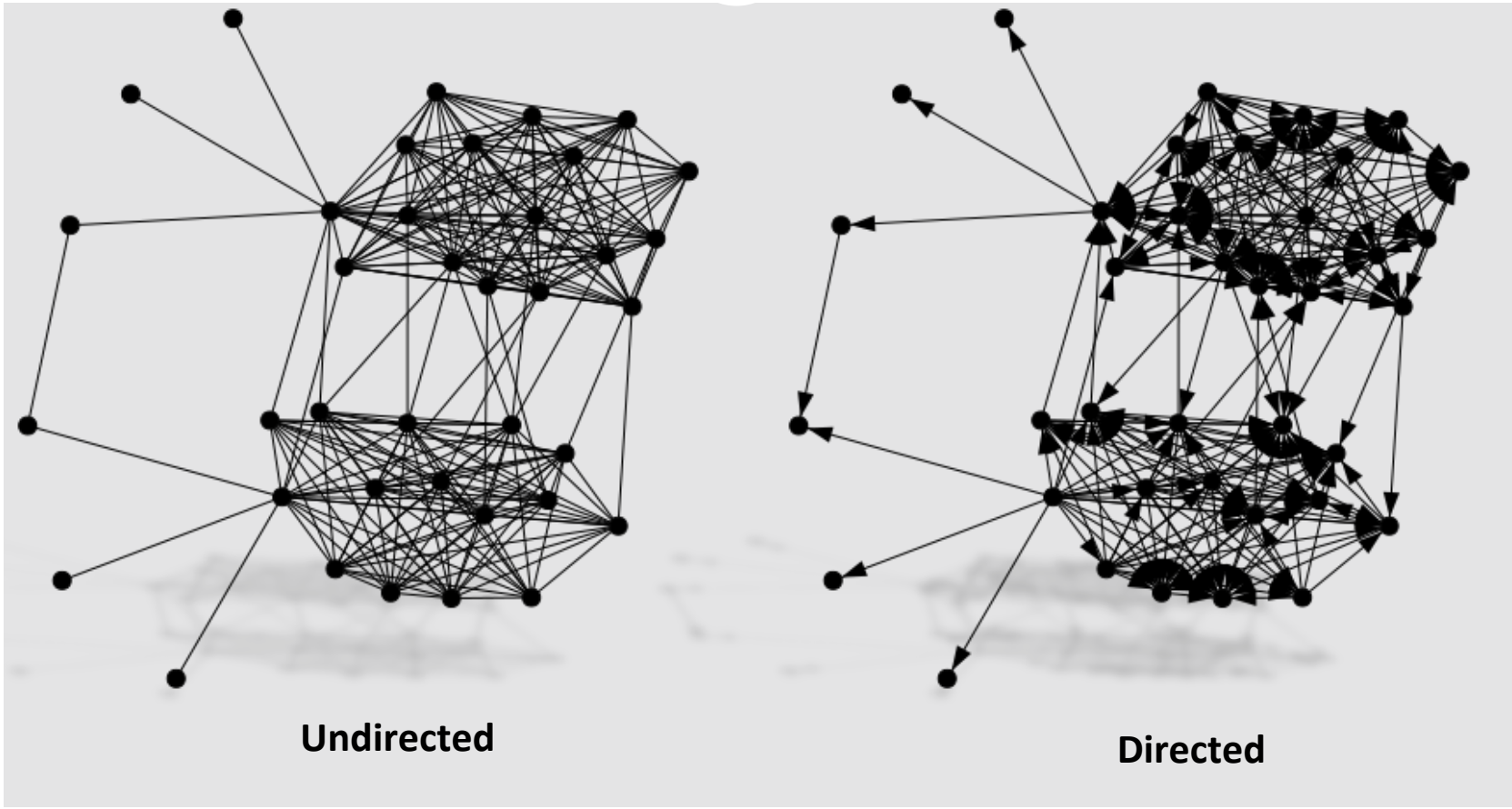
they arise out of an affiliation network! the one-mode projection we observe is an embedding of a multidimensional network that exists.

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



# practical aspects



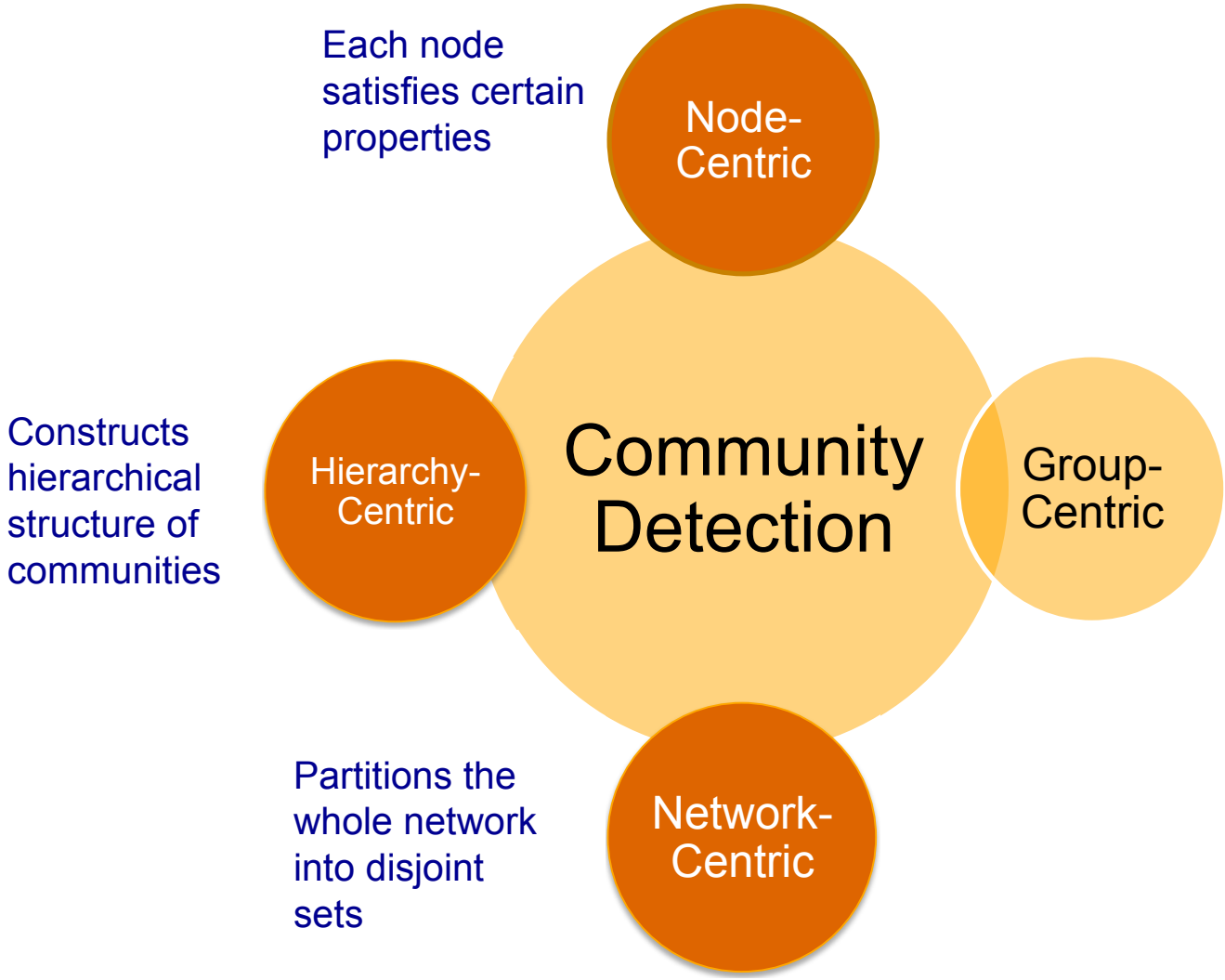

Many methods:

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

# Cohesive Subgroups: A Typology

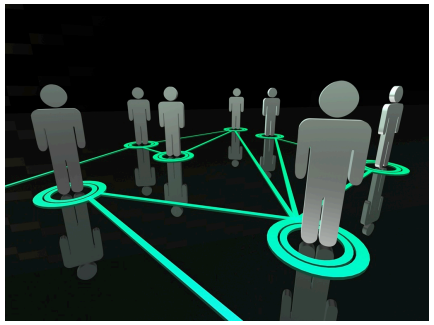
| | Found by algorithm (**input** data driven) | Found by finding sets with **output** properties |
|---|---|---|
| Network / Graph theory | *Graph-theoretic data driven algorithms*<br>Newman-Girvan | *Formal definitions of sociological groups*<br>*{mathematical ethnography}*<br>Clique, n-clique, n-clan,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

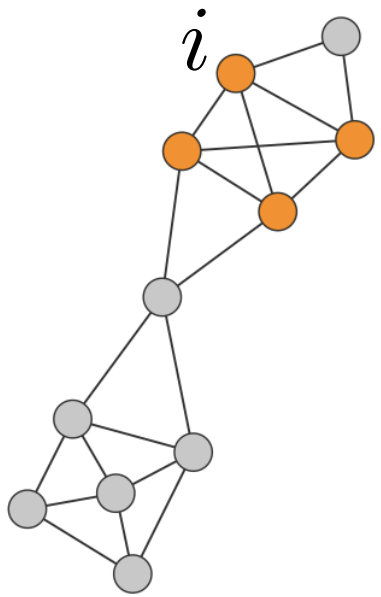
*Internal degree,* $k_i^{int}$ *:* set of links of node *i* that connects to other nodes of the same community *C*.

*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.

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



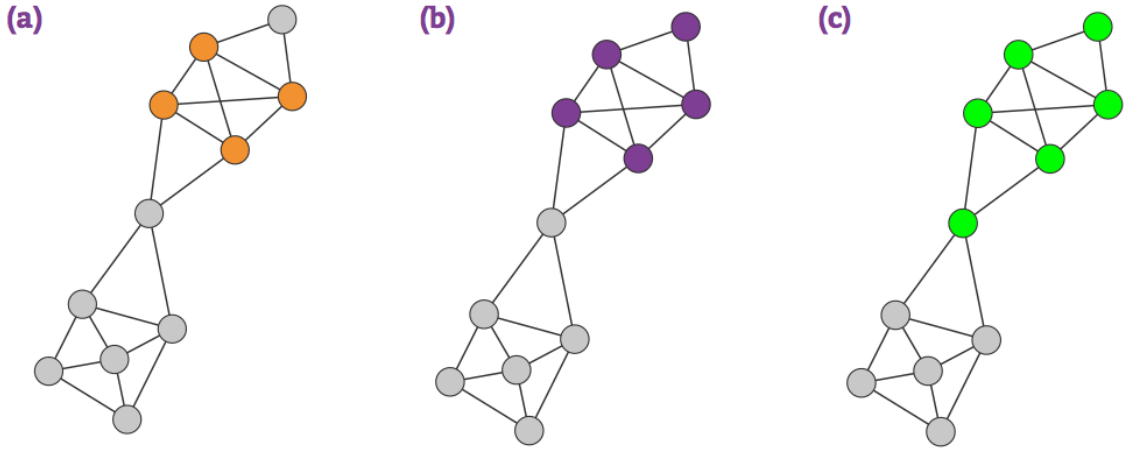
# Fundamental Hypotheses of communities

*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:

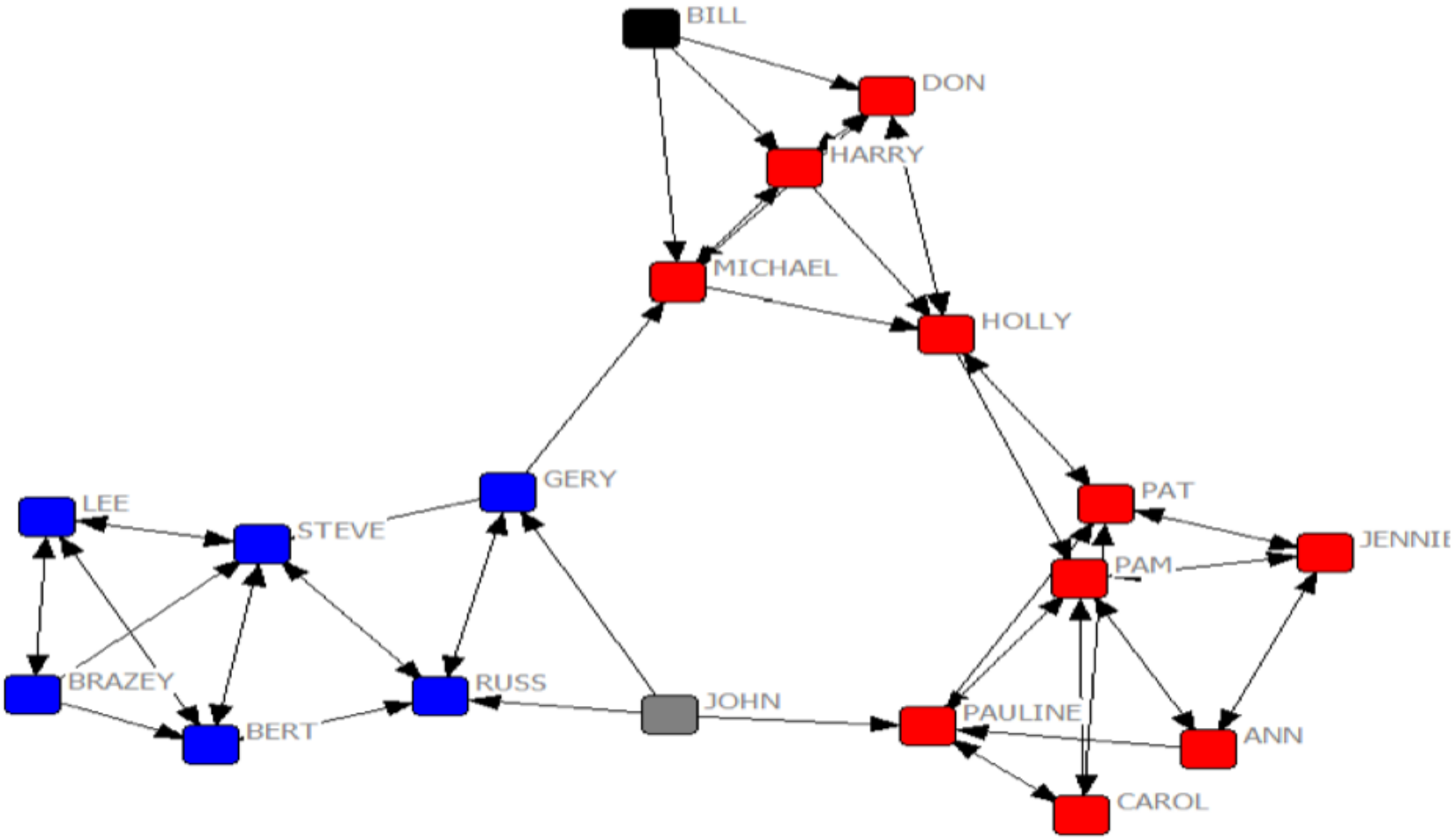
- **Complete Mutuality** [everybody in the group knows everybody else]
  - components
  - cliques
- **Reachability of members** [individuals are separated by at most n hops]
  - n-clique, n-clan, n-club
- **Nodal degrees** [everybody in the group has links to at least *k* others in the group]
  - k-plex, 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

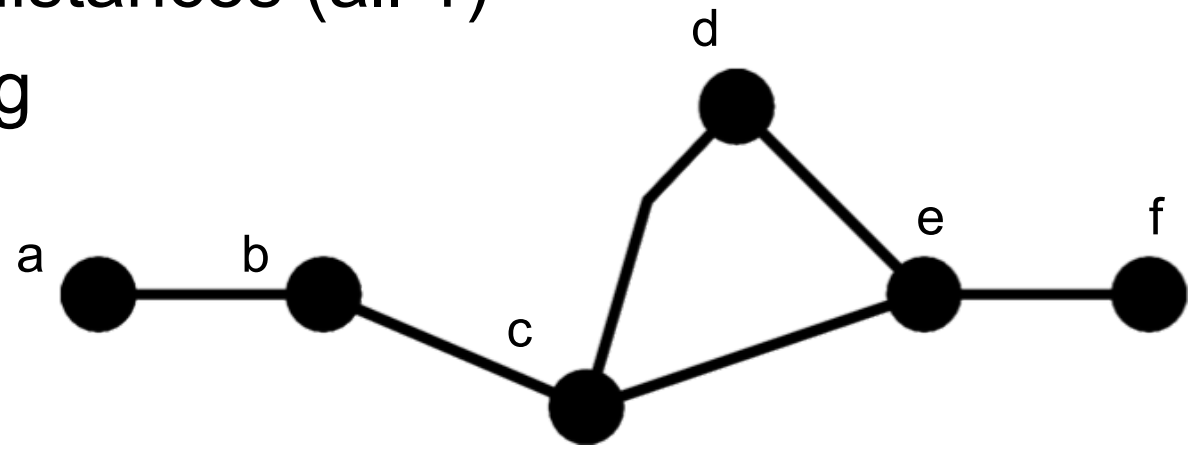
## Colored by Strong Components



# complete mutuality | **cliques**

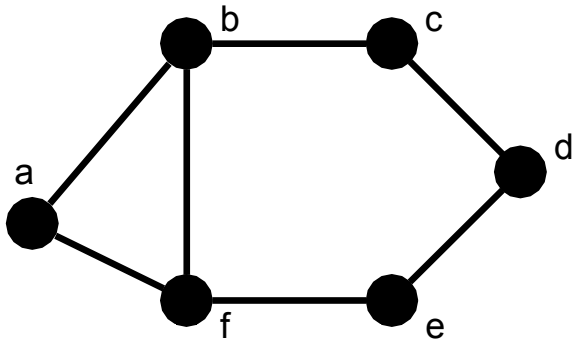
- Definition
  - Maximal, complete subgraph
  - Set S s.t. for all u,v in S, (u,v) in E
- Properties
  - Maximum density (1.0)
  - Minimum distances (all 1)
  - overlapping
  - Strict

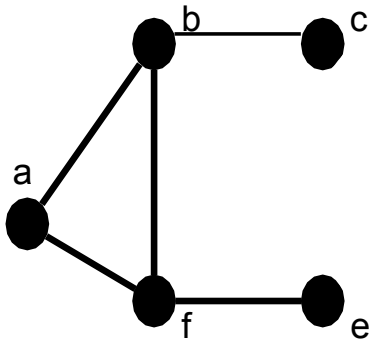
{c,d,e} is the only clique



# 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

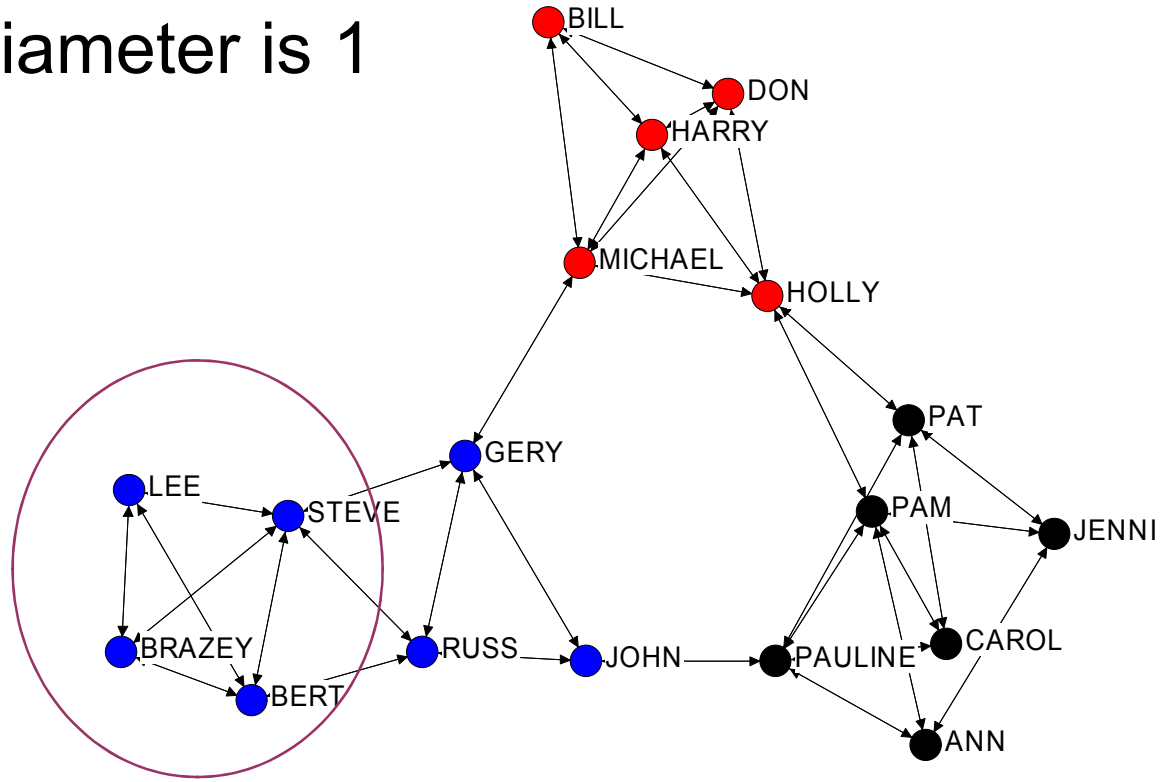



Subgraph induced by {a,b,c,f,e}

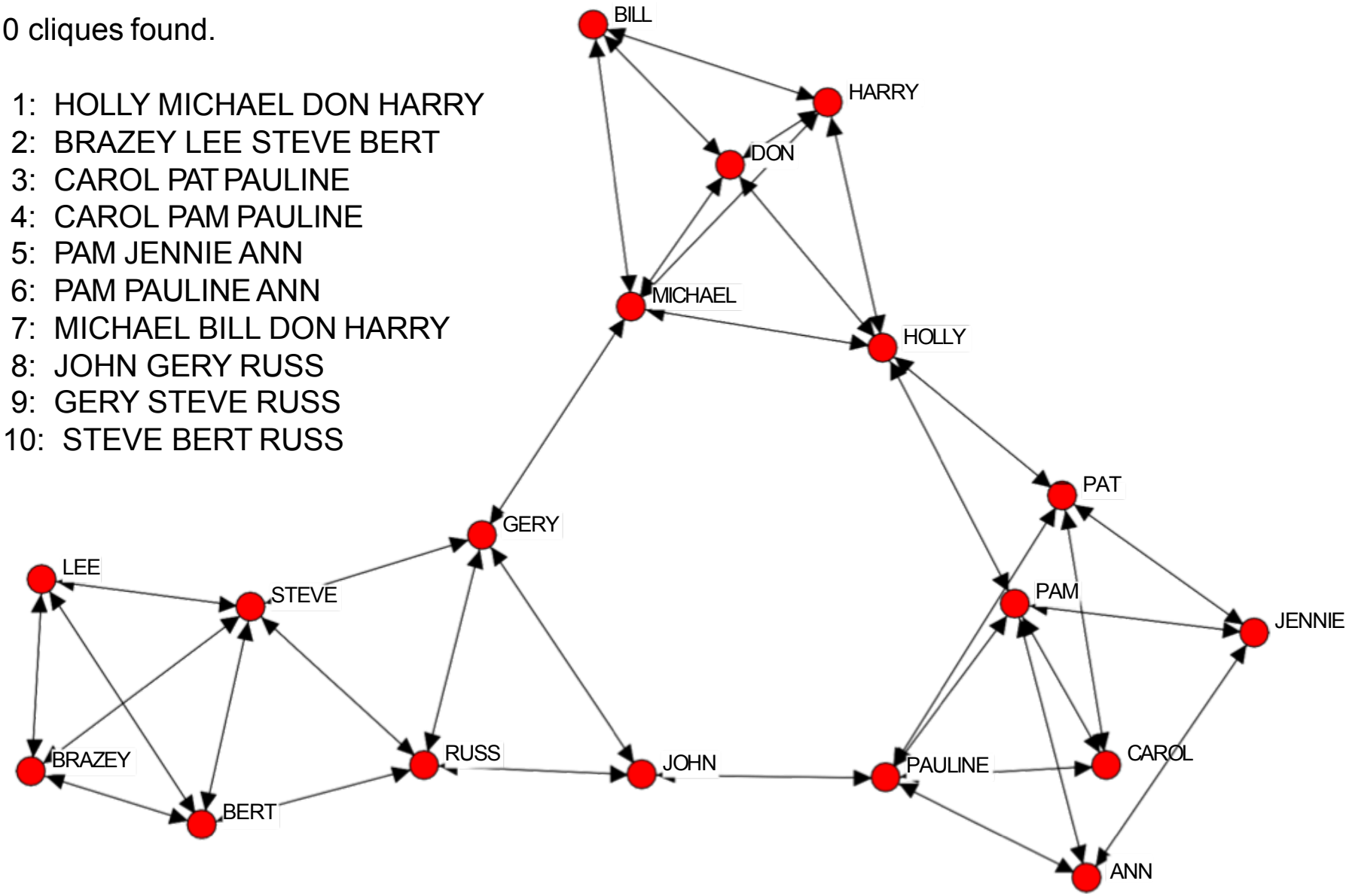
# complete mutuality | clique

- 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

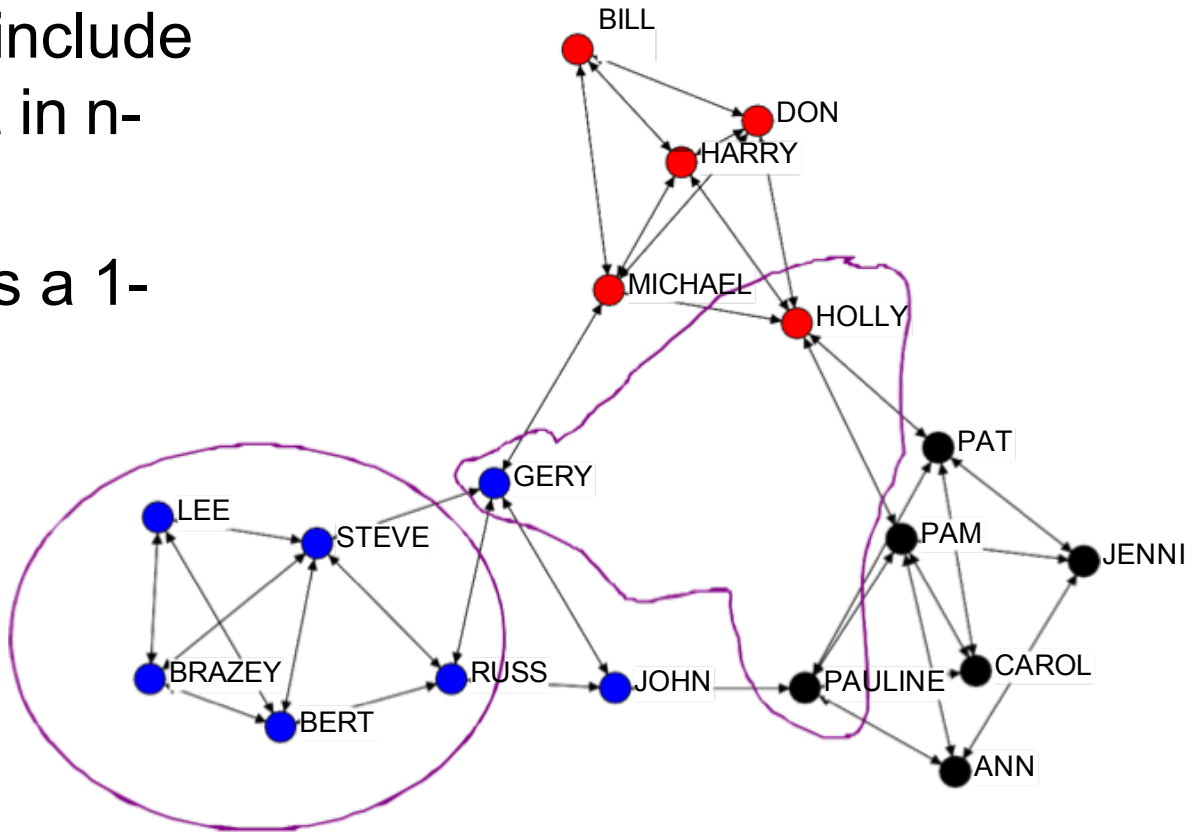
- Distance Relaxations (length of paths)
  - n-clique
  - n-clan
  - n-club
- Density Relaxations (number of ties)
  - k-plex
  - k-core

# 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

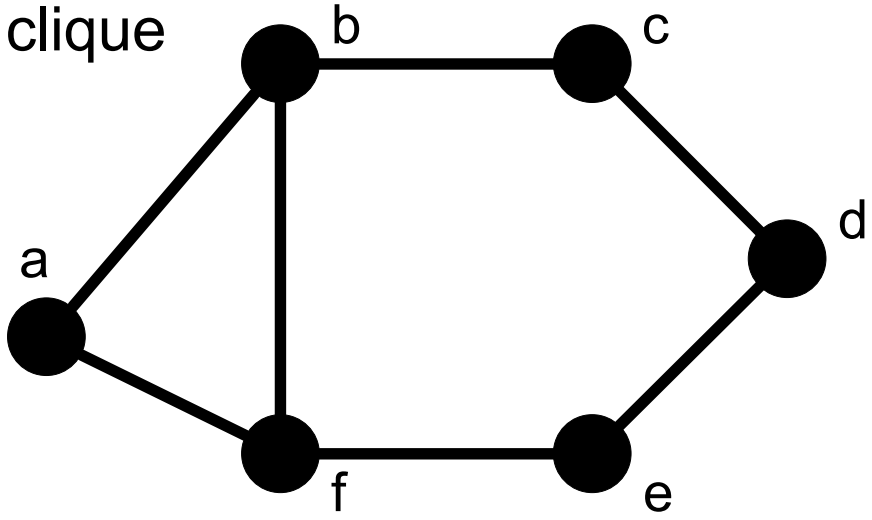
Is this a 2-Clique?
NO!
What about now?

But so is this!!!



# reachability of members | n-clique

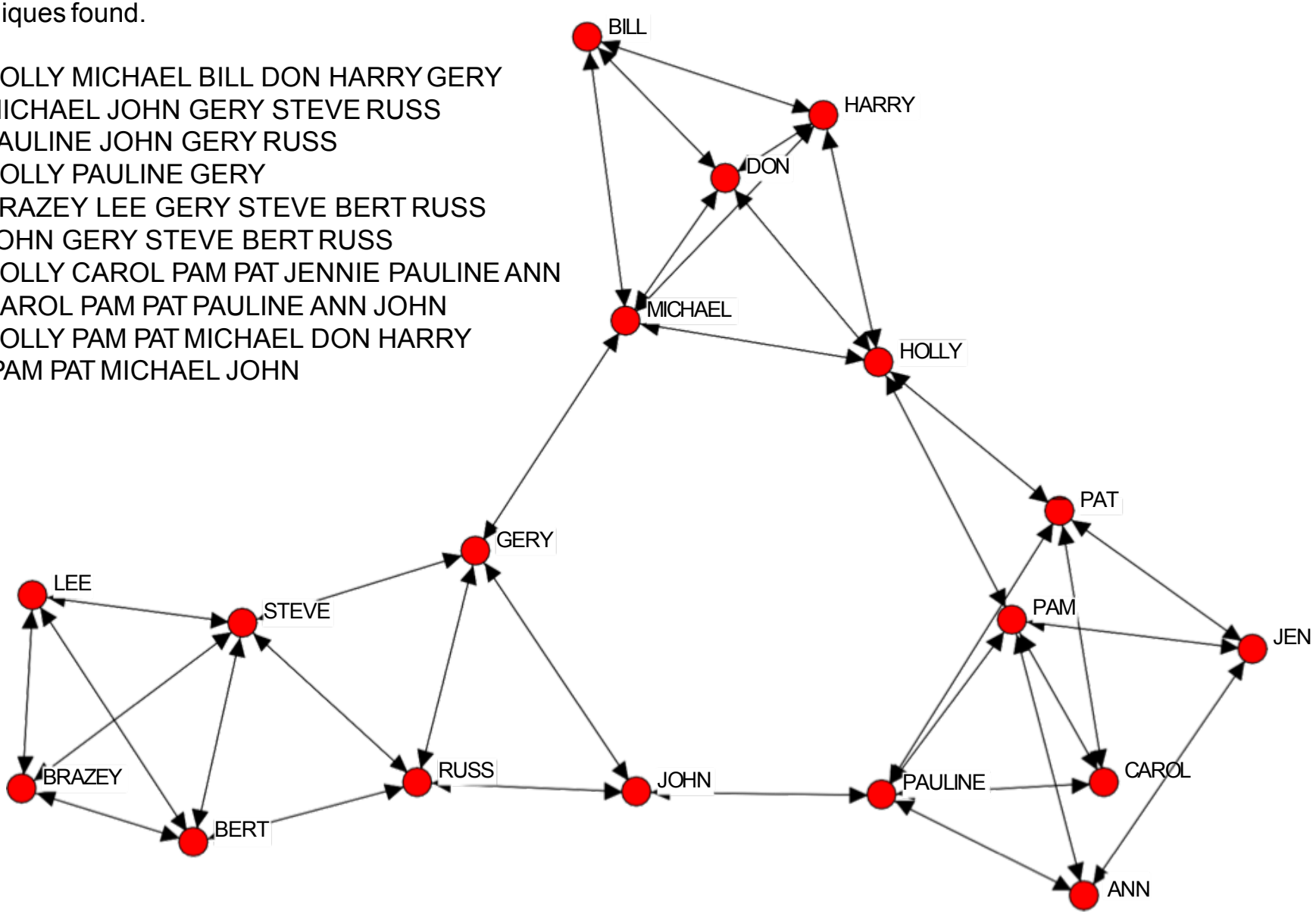
- Definition
  - Maximal subset s.t. for all *u,v* in S, d(*u,v*) <= n
  - Distance among members less than specified maximum
  - When n = 1, we have a clique
- Properties
  - Relaxes notion of clique
    - Avg distance can be greater than 1



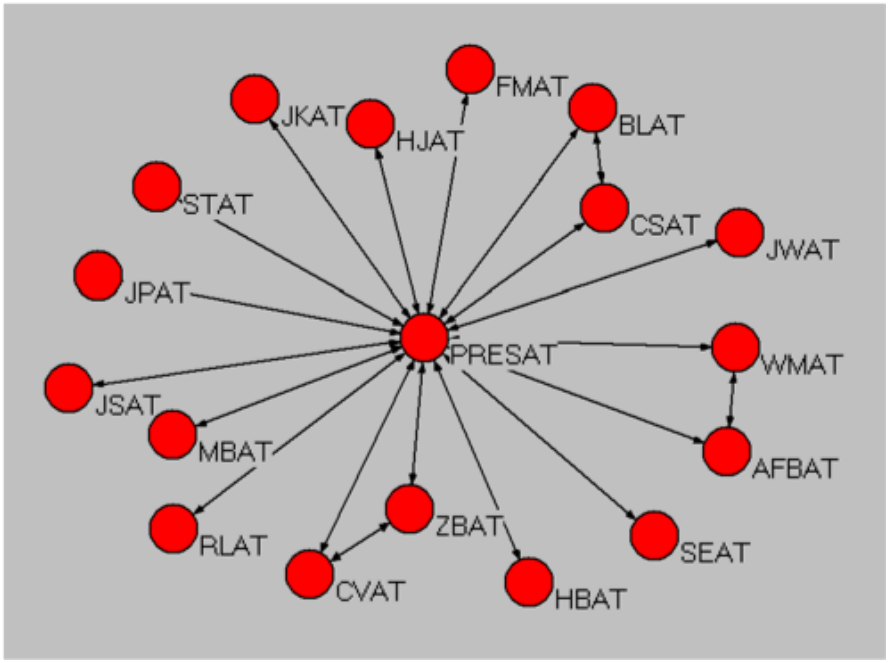

Is {a,b,c,f,e} a 2-clique?
yes

10 2-cliques found.

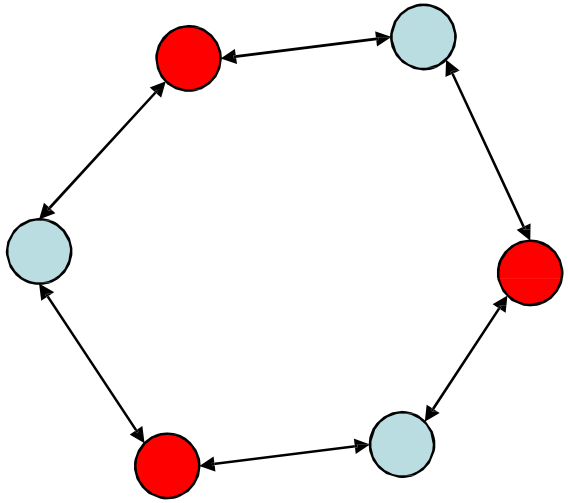
1: HOLLY MICHAEL BILL DON HARRY GERY
2: MICHAEL JOHN GERY STEVE RUSS
3: PAULINE JOHN GERY RUSS
4: HOLLY PAULINE GERY
5: BRAZEY LEE GERY STEVE BERT RUSS
6: JOHN GERY STEVE BERT RUSS
7: HOLLY CAROL PAM PAT JENNIE PAULINE ANN
8: CAROL PAM PAT PAULINE ANN JOHN
9: HOLLY PAM PAT MICHAEL DON HARRY
10: PAM PAT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



# Some are counter-intuitive (And not necessarily cohesive)



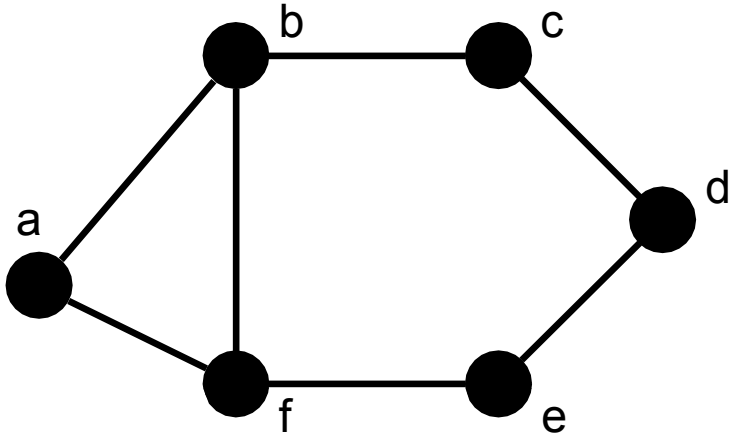

This is a 2-Clique

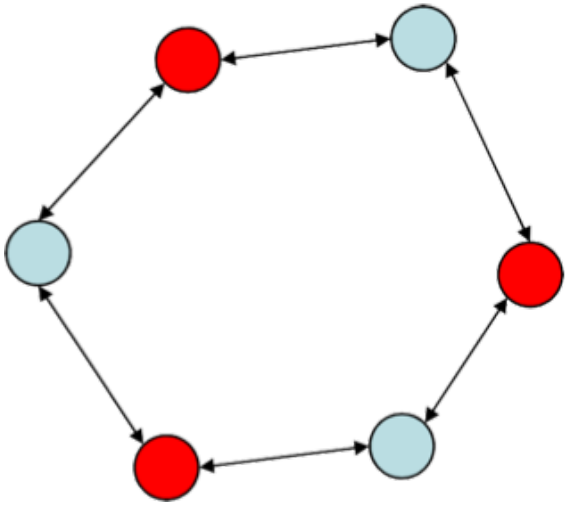


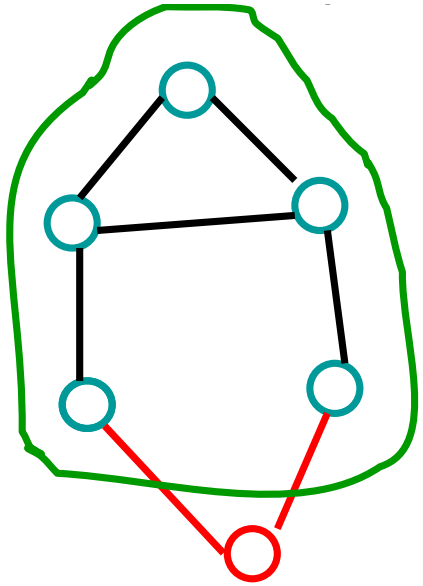
Red Nodes form a 2-Clique, so do Blues

# 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



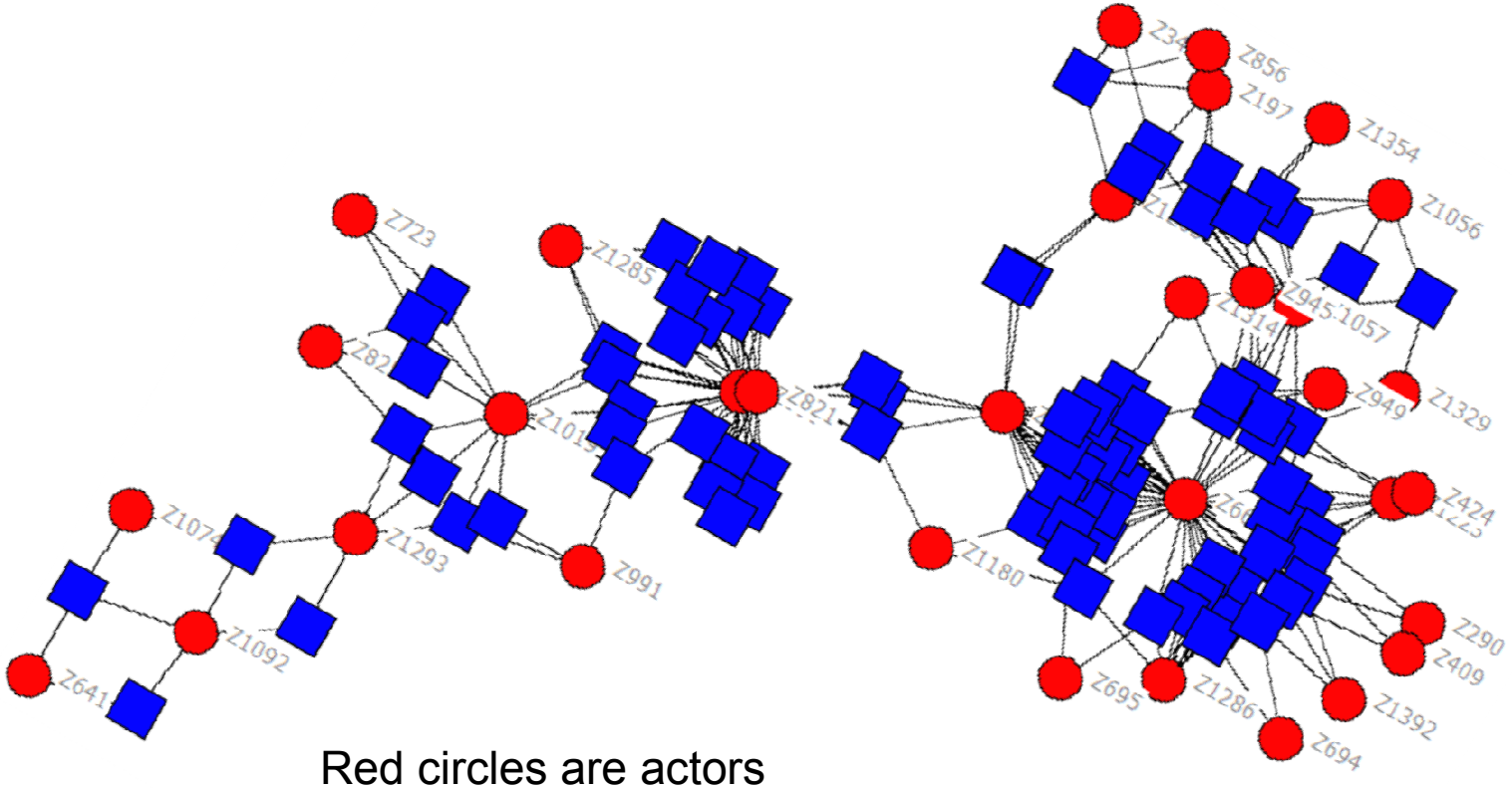


2 – clique
diameter = 3

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

# 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**

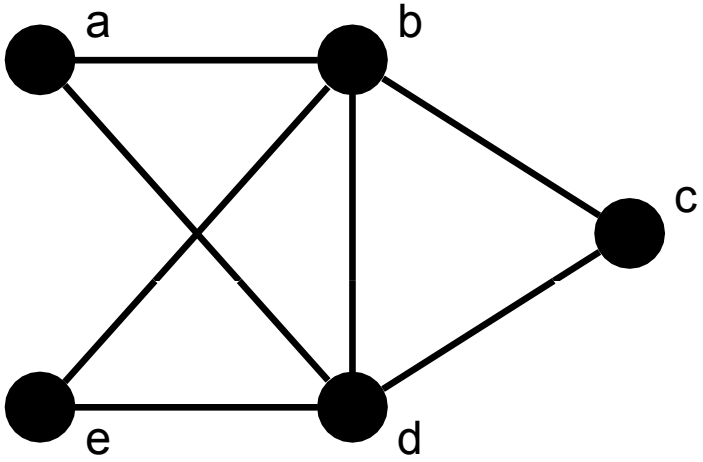
- But, Cliques also have **maximum density** (d = 1), and we can relax that definition instead.

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

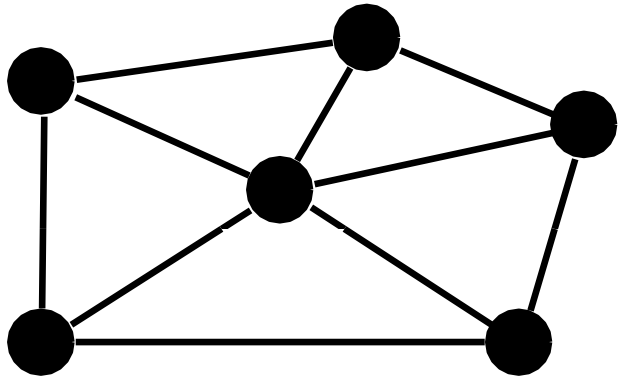
# nodal degrees | **k-plex**

- 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?

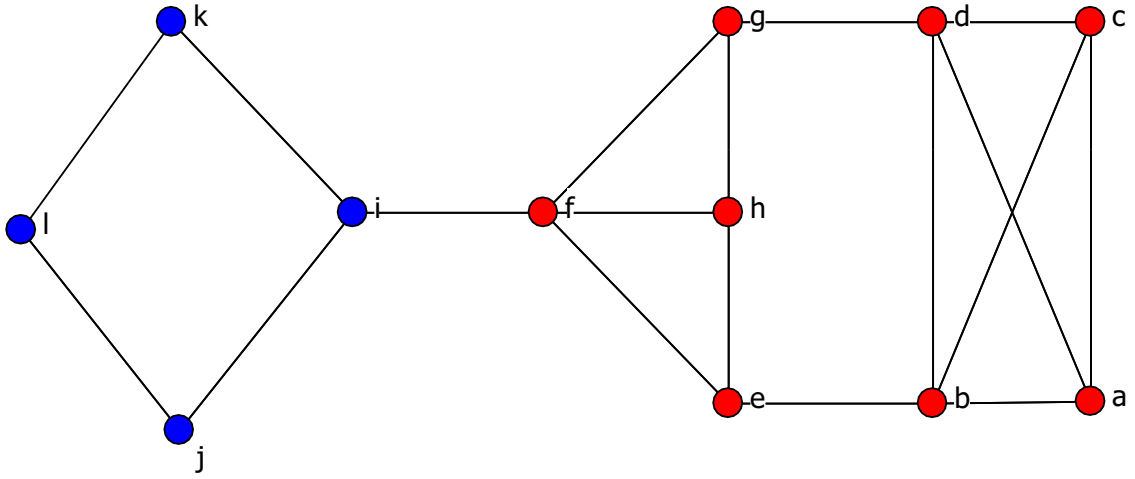
# nodal degrees | **k-core**

- 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

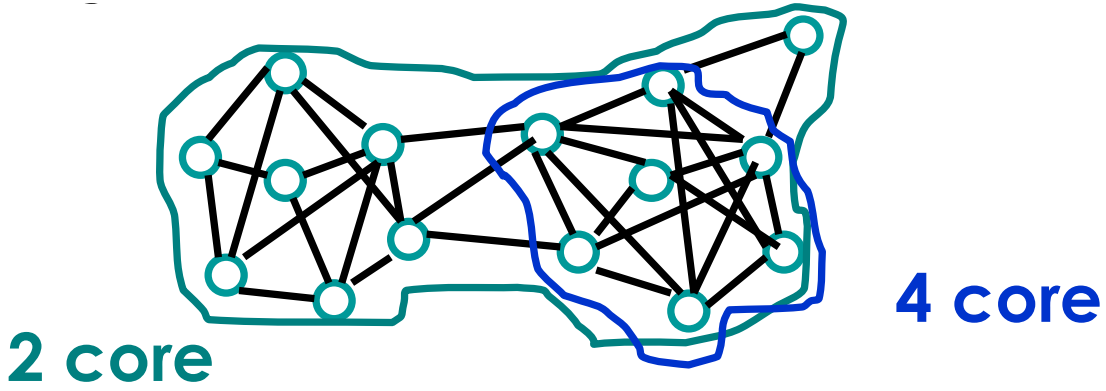
# formal definition

- A k-core is a maximal subgraph such that for all u in S, $\alpha$(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…

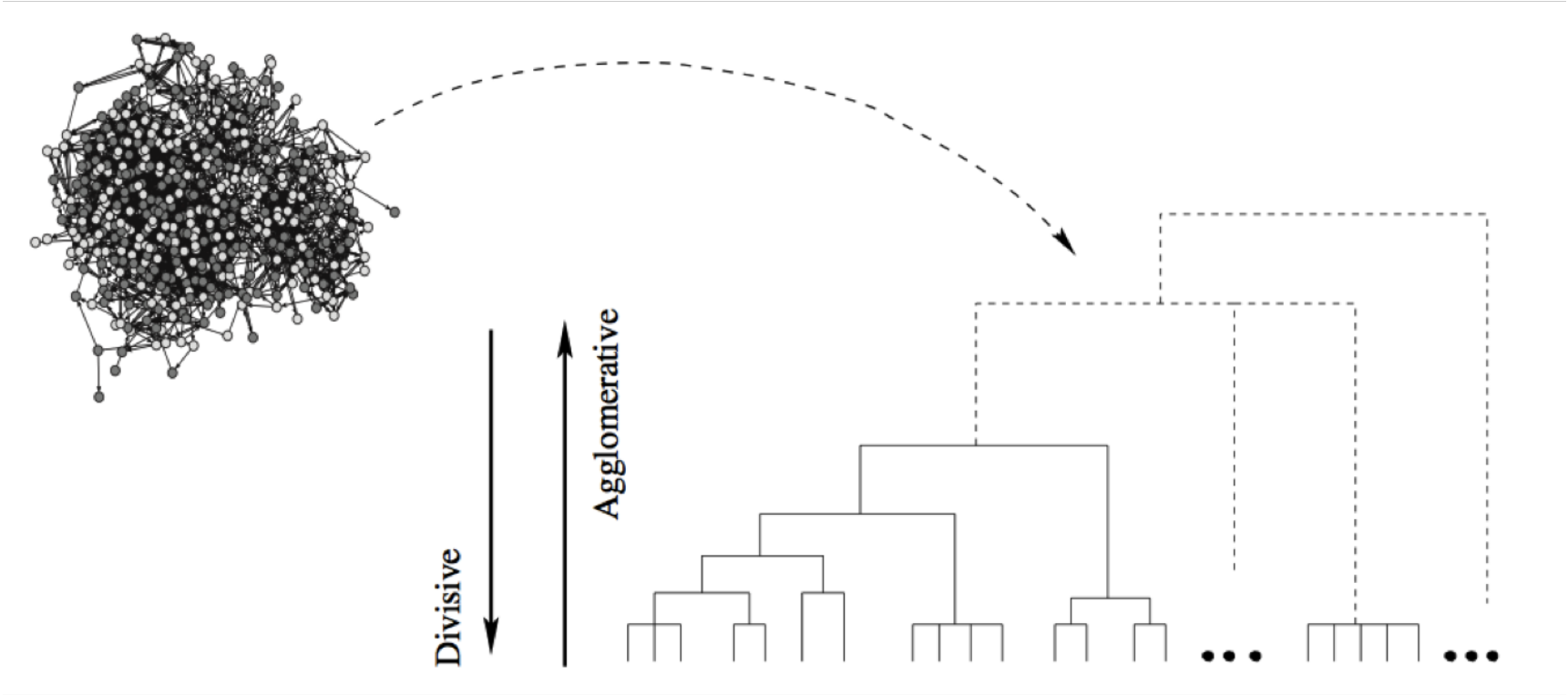
# recap node-centric communities (cohesive subgroups)

- 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:

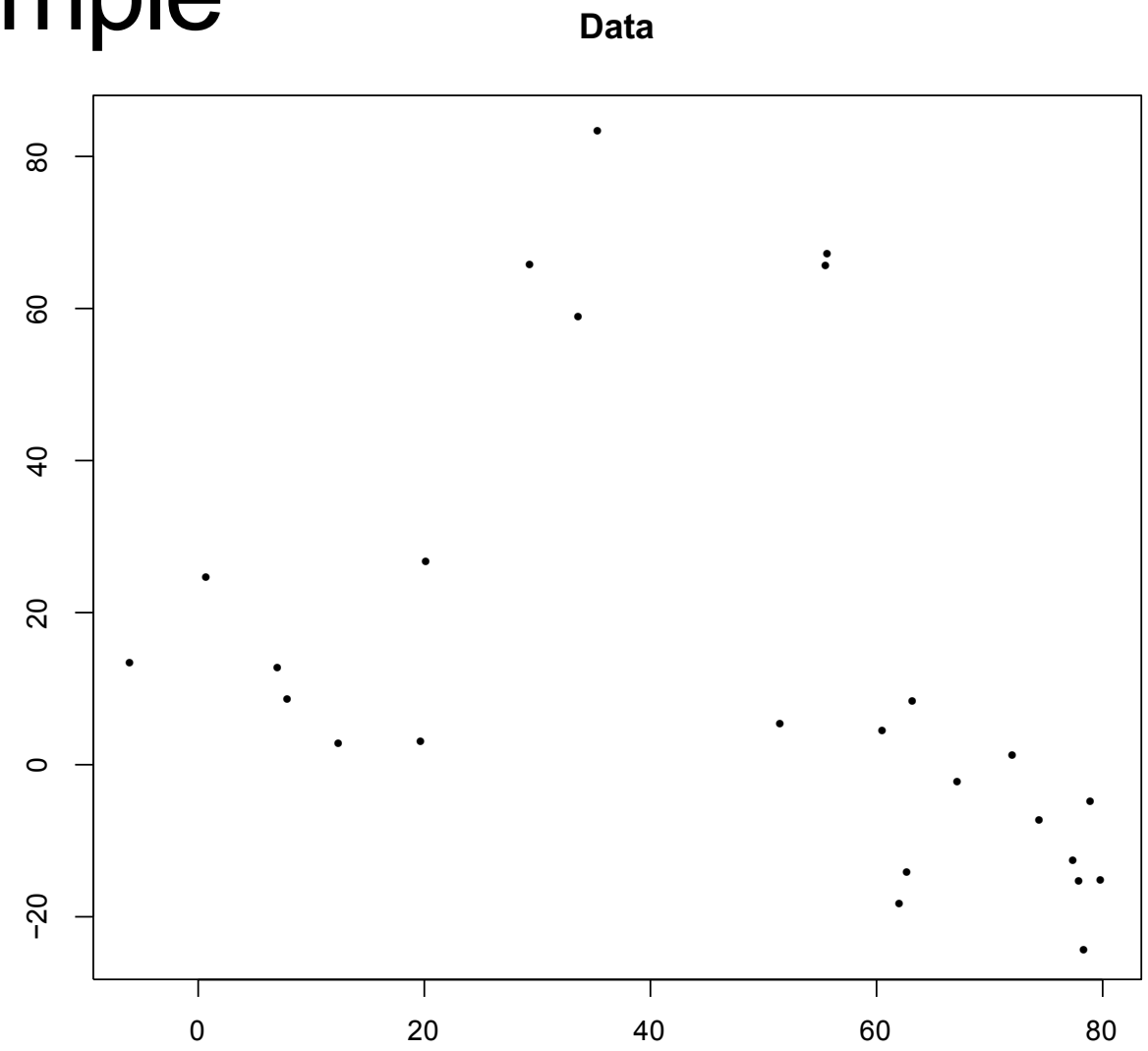
- Define similarity measure between vertices based on network structure
  - Jaccard similarity
  - Cosine similarity
  - Pearson correlation
  - Eucledian distance (dissimilarity)
- Calculate similarity between all pairs of vertices in the graph (similarity matrix)
- Group together vertices with high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

## Pseudocode

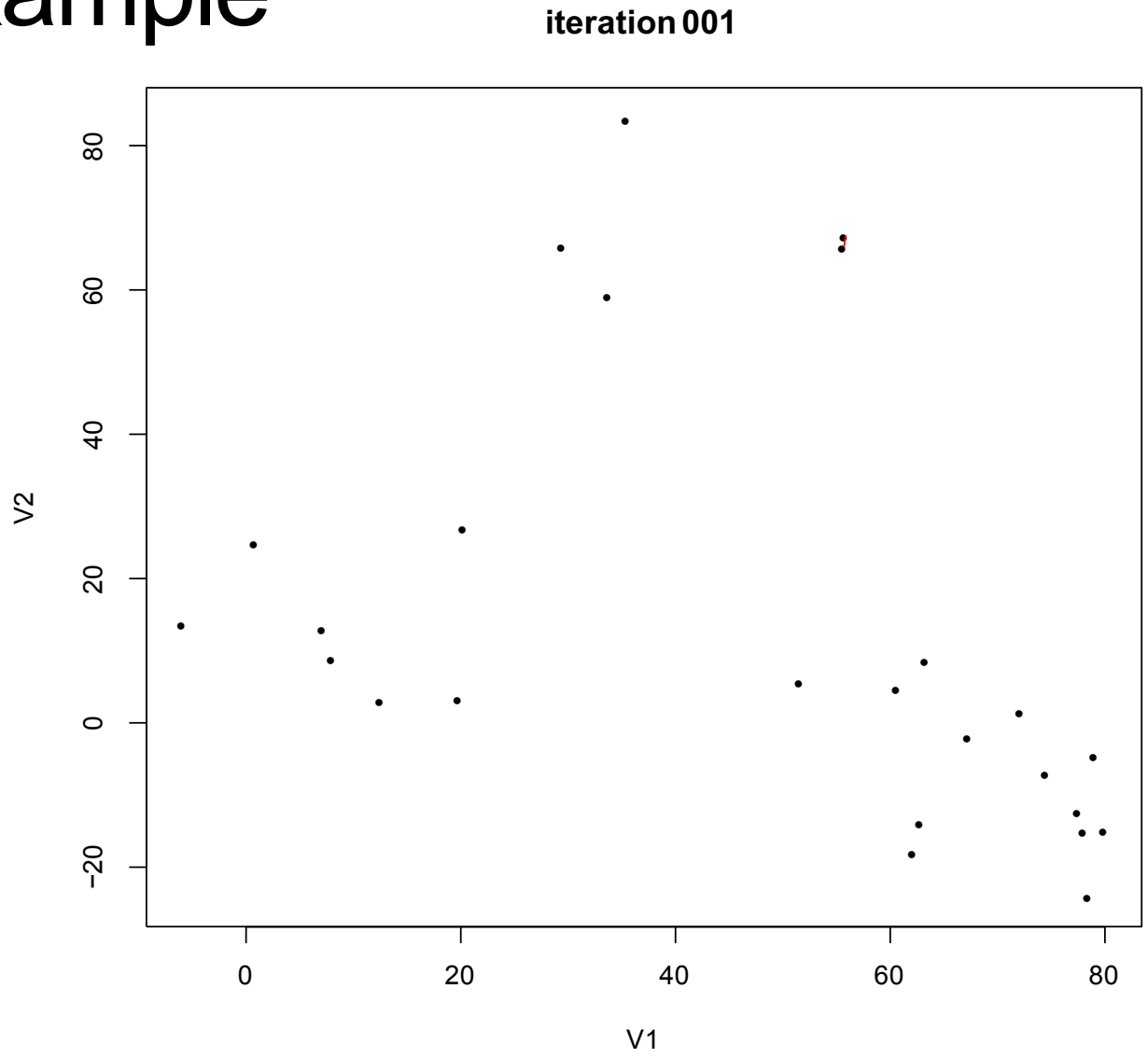
1. Assign each node to its own cluster
2. Find the cluster pair with highest similarity and join them together into a cluster
3. Compute new similarities between new joined cluster and others
4. Go to step 2 until all nodes form a single cluster

# Network-Centric | [Agglomerative] Community Detection

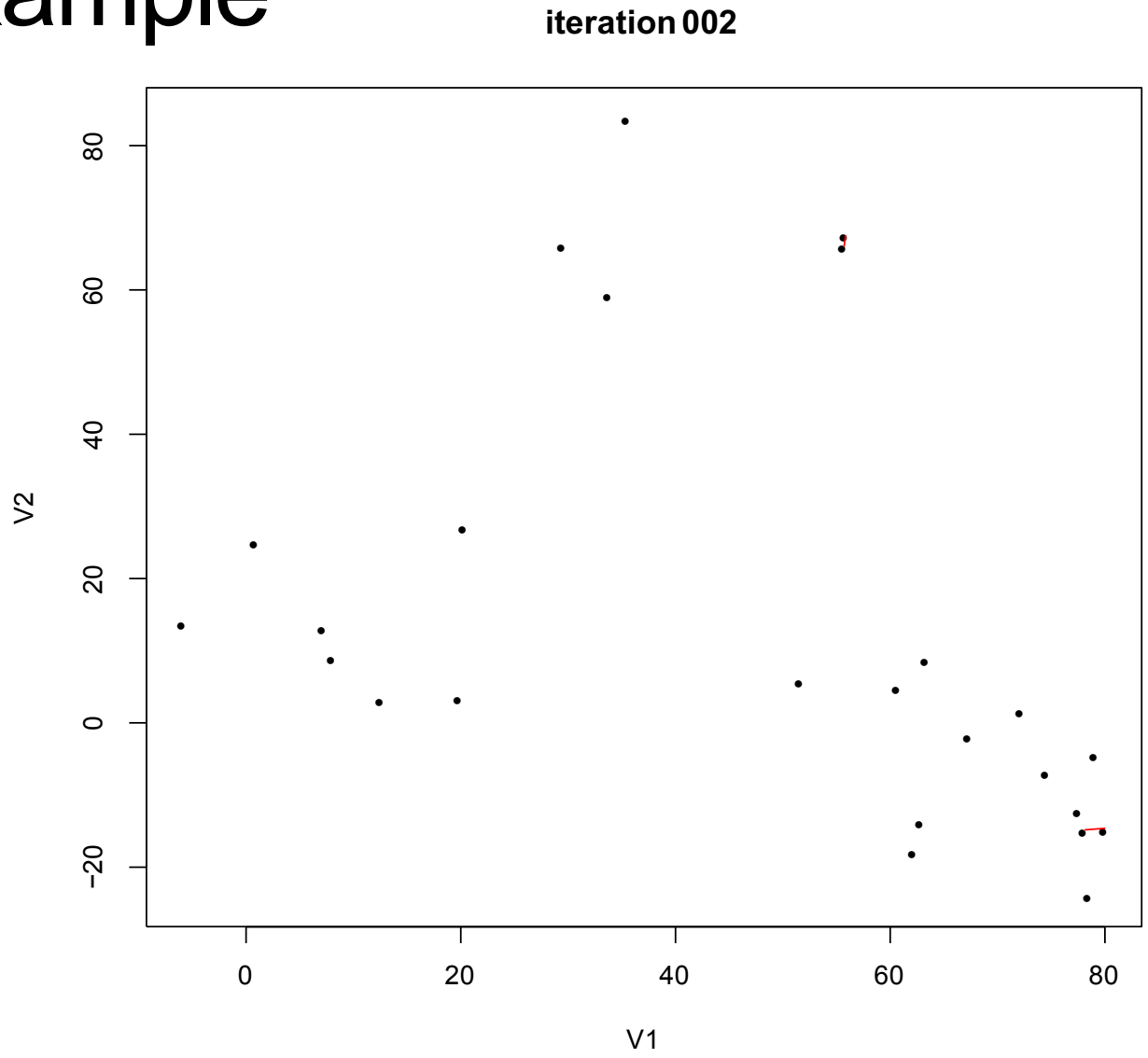
## Example



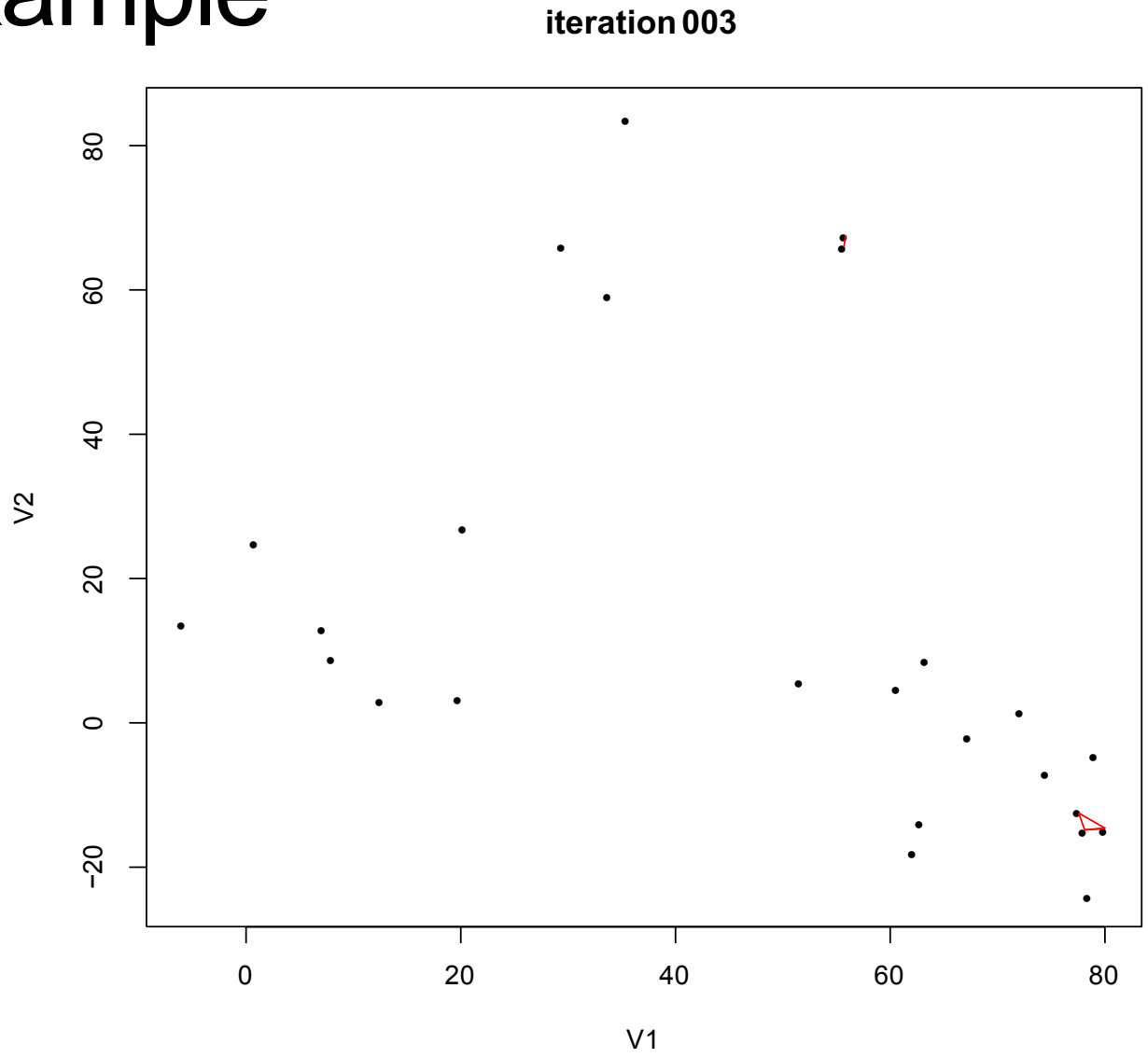
# Example



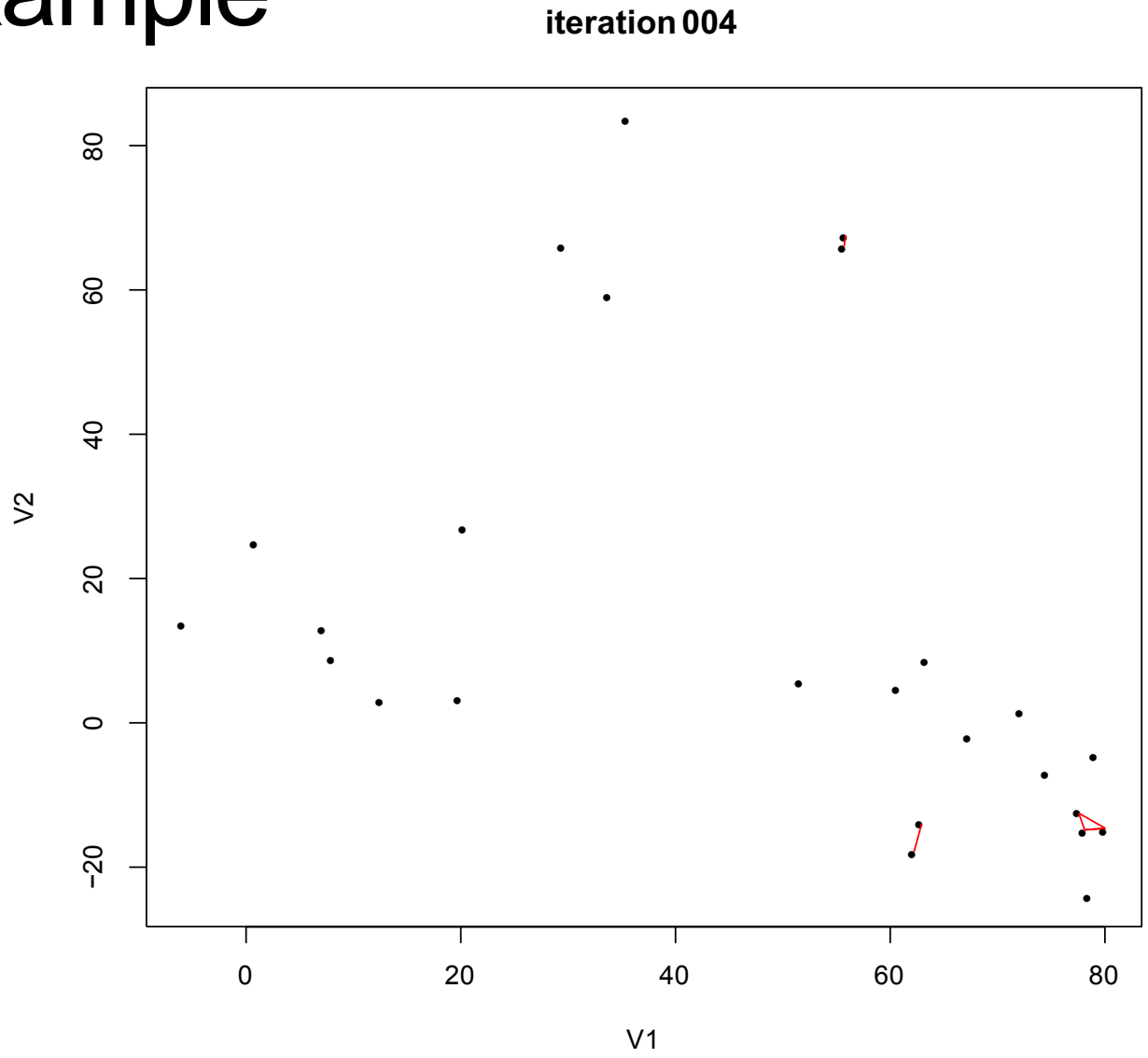
# Example



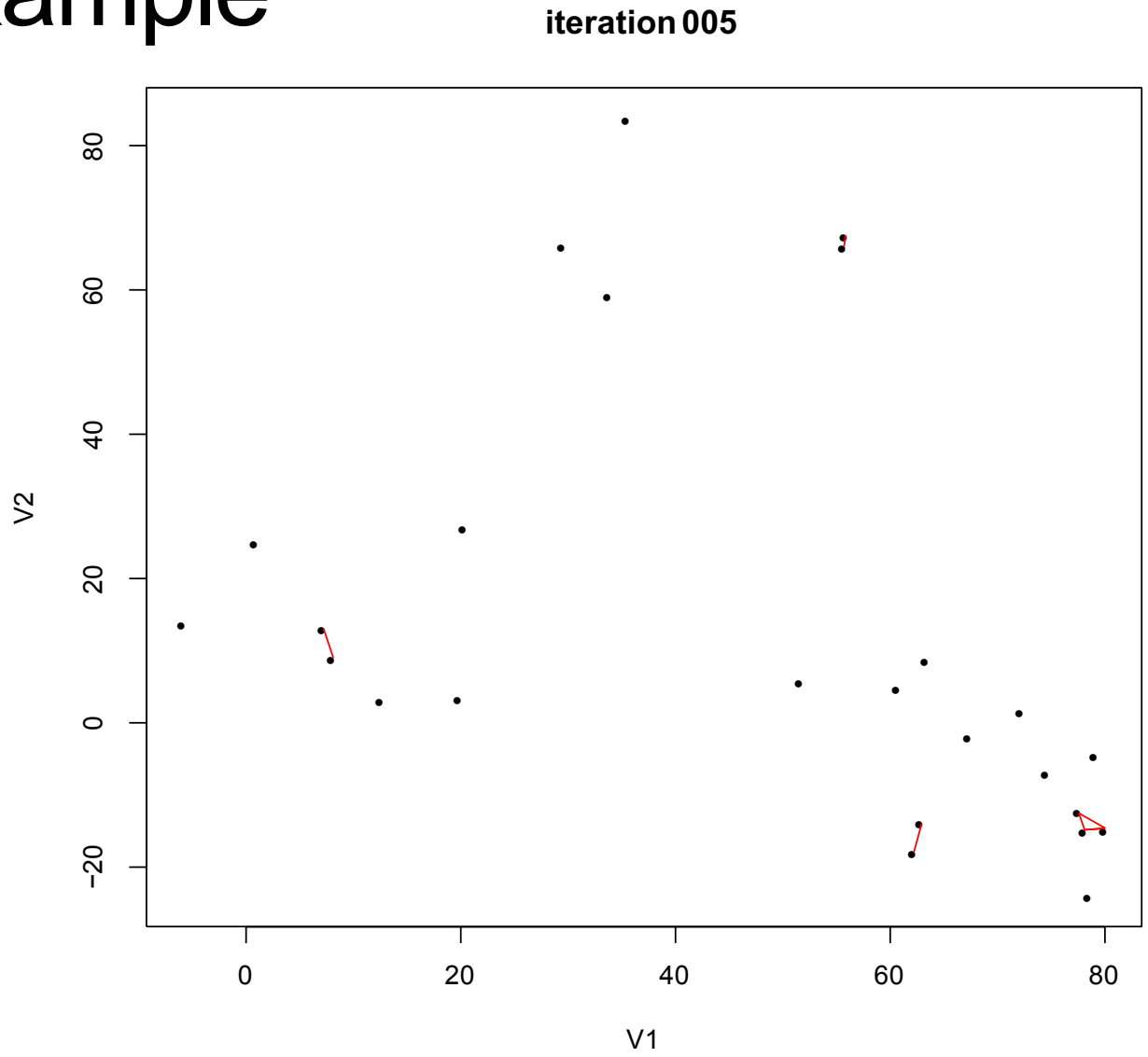
# Example



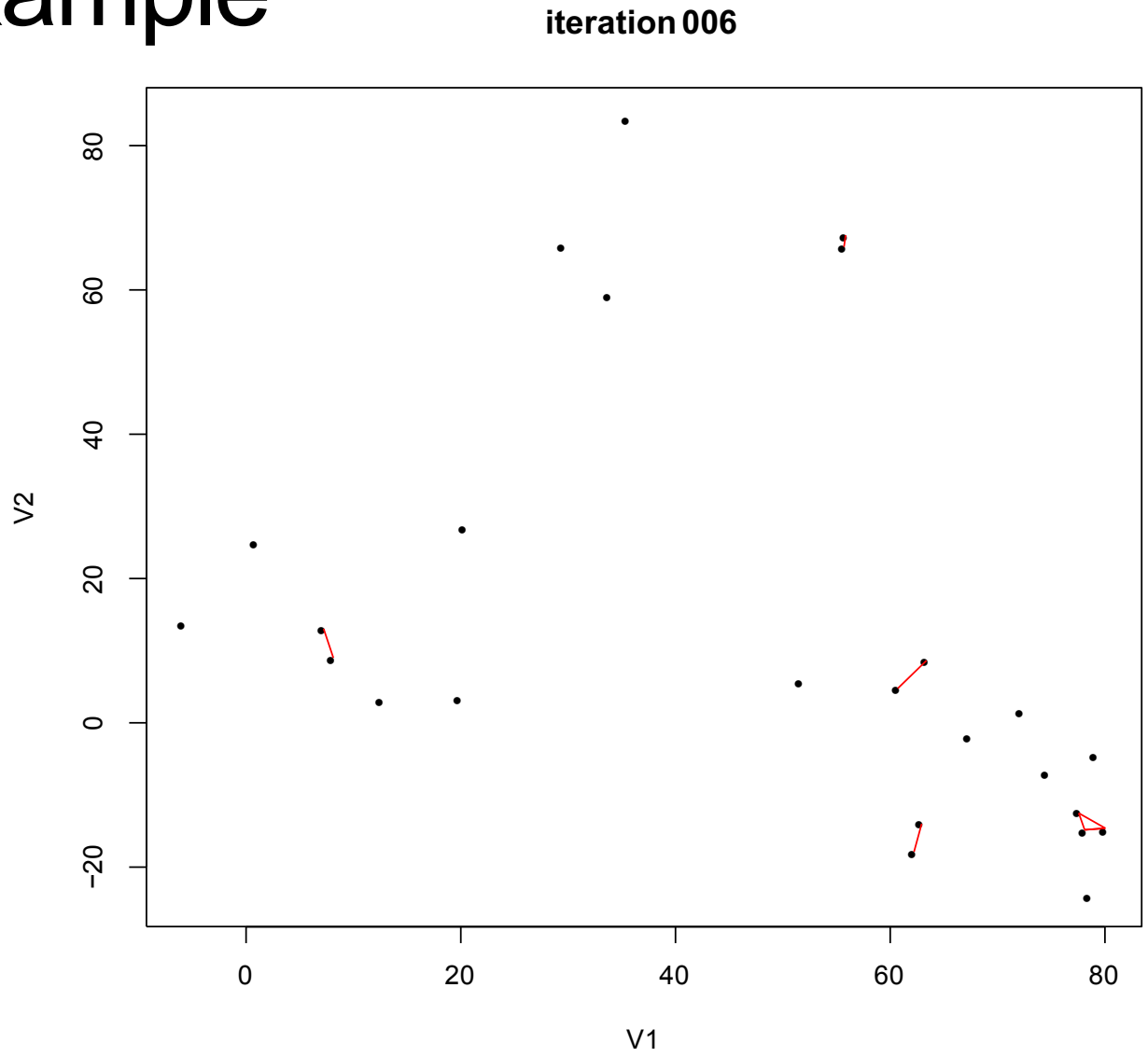
# Example



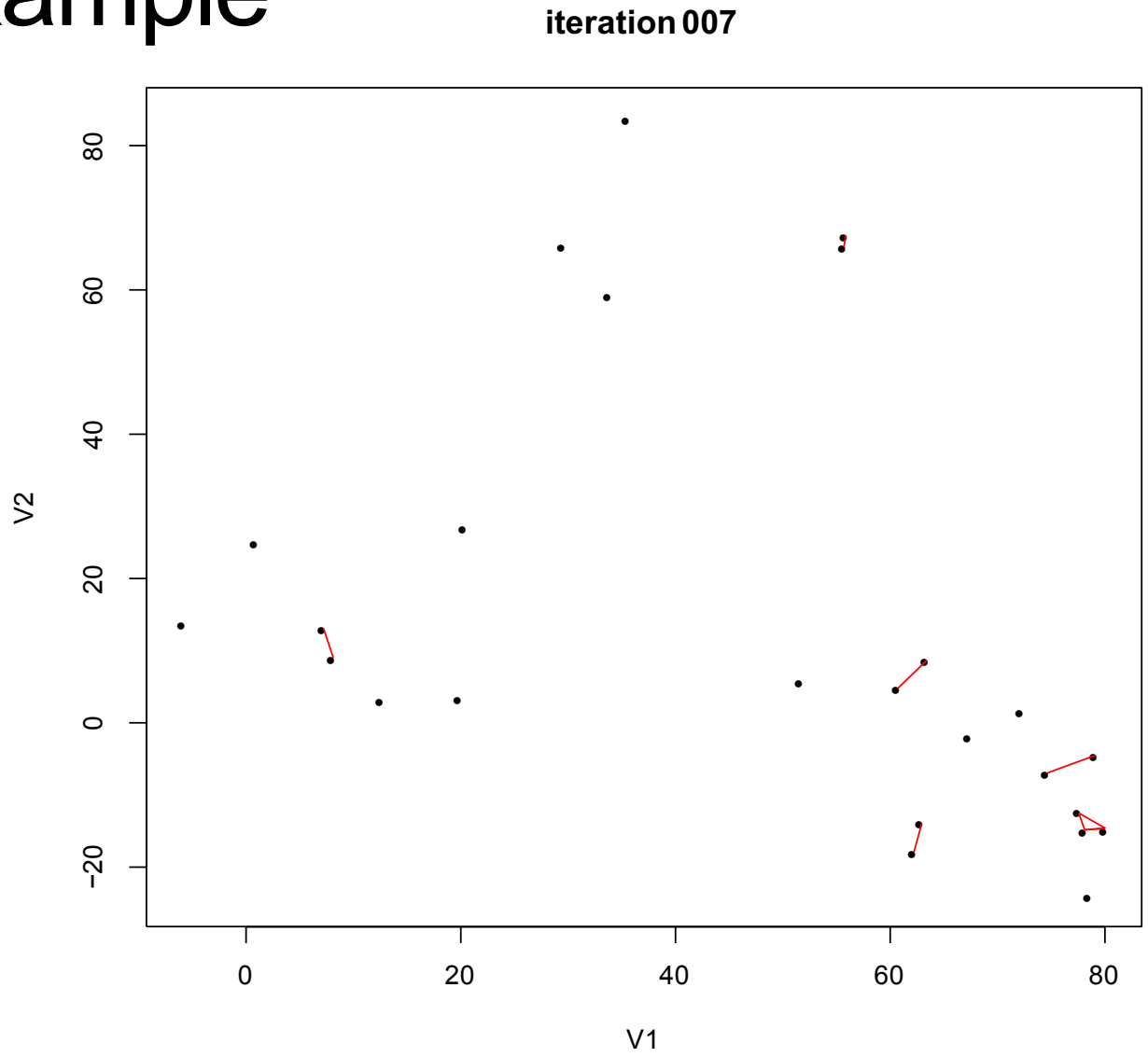
# Example



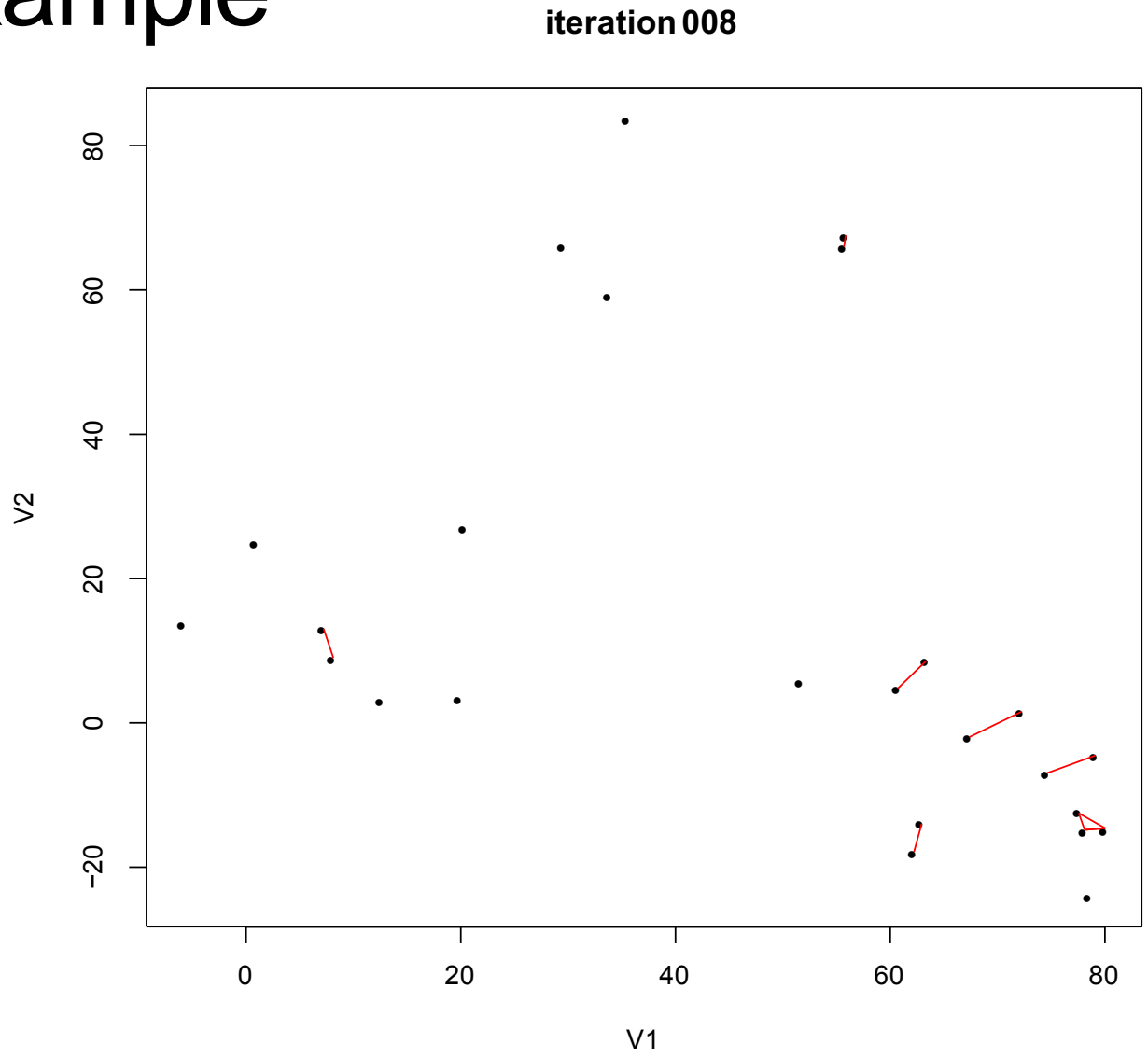
# Example



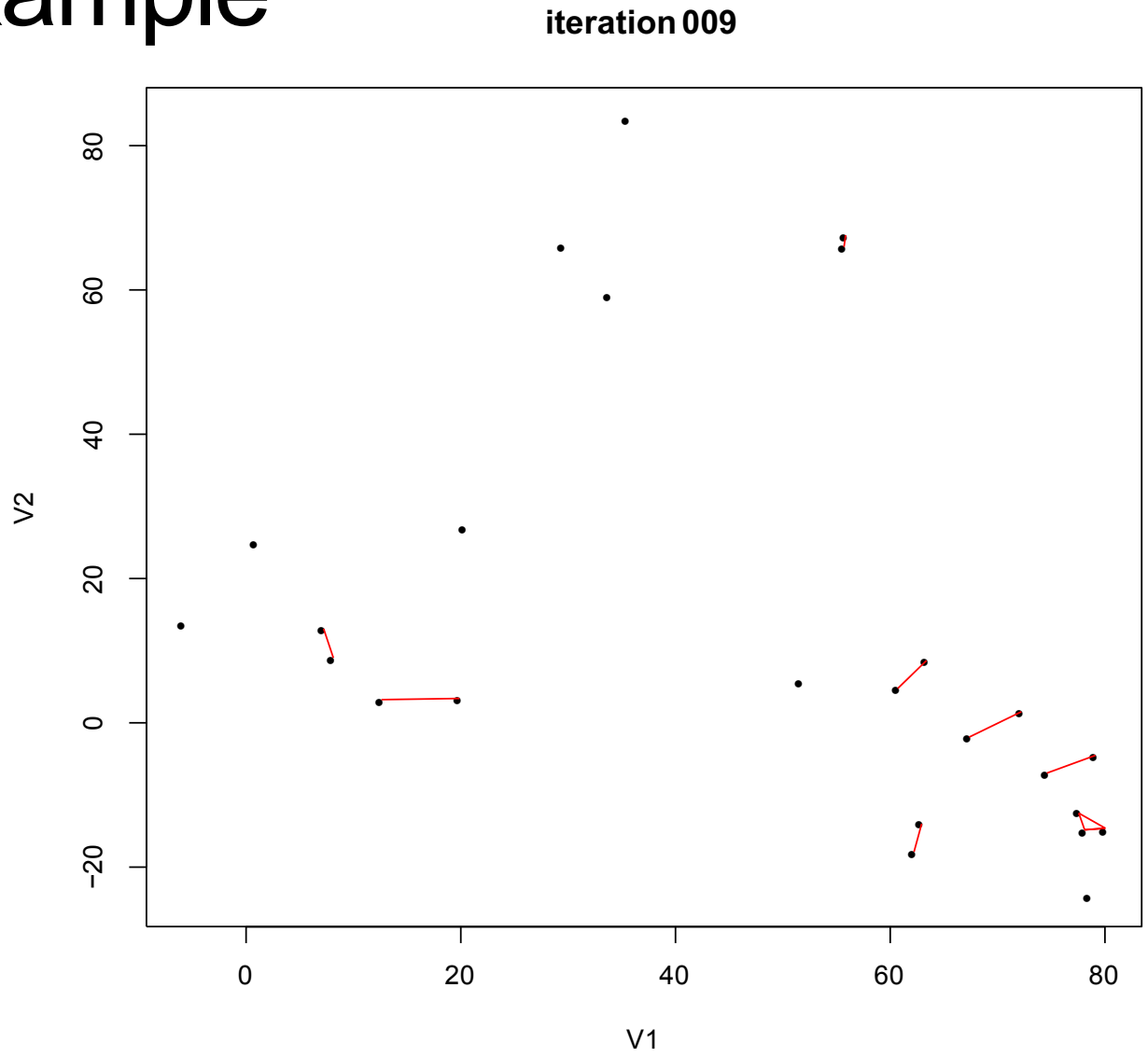
# Example



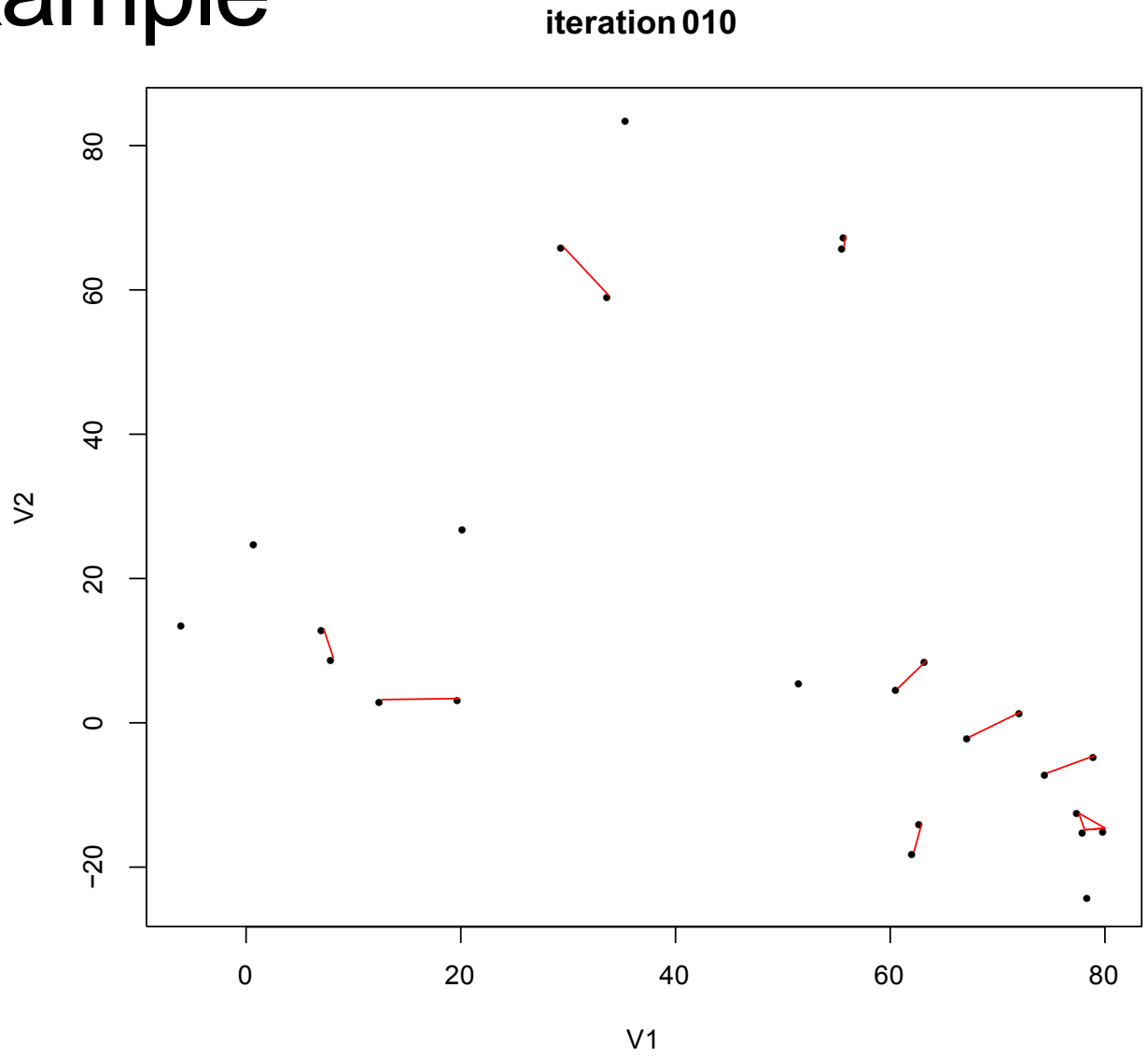
# Example



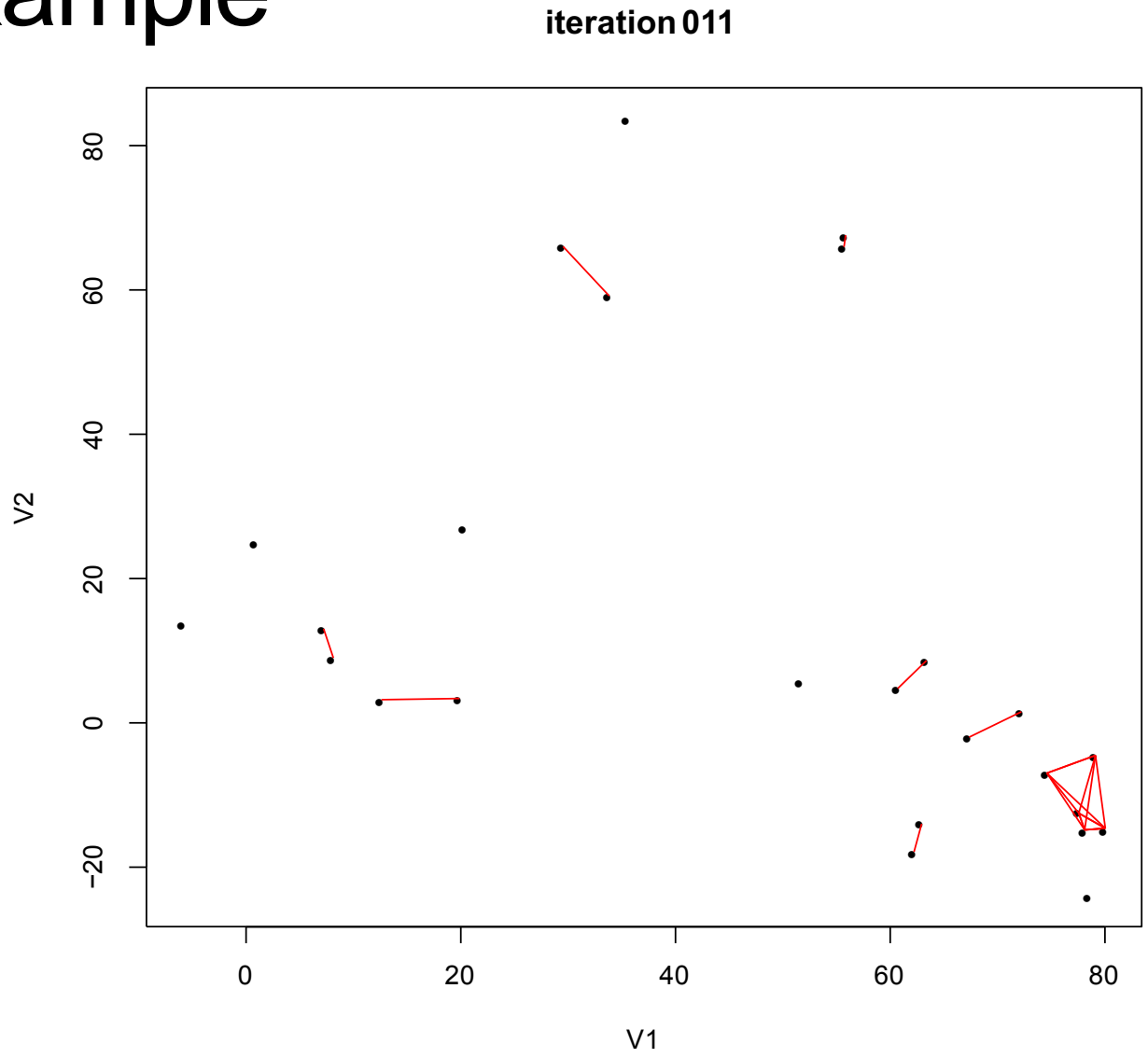
# Example



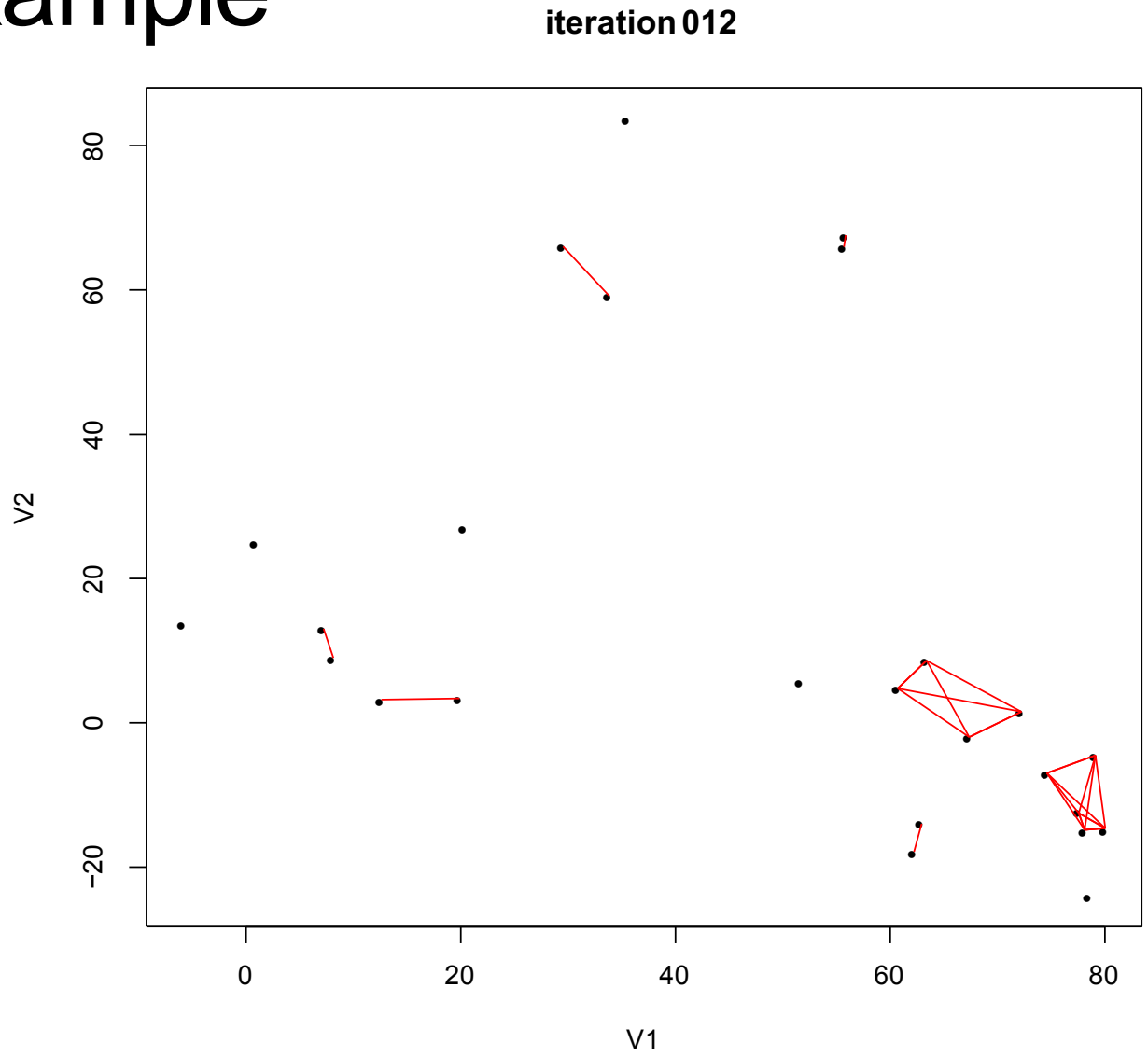
# Example



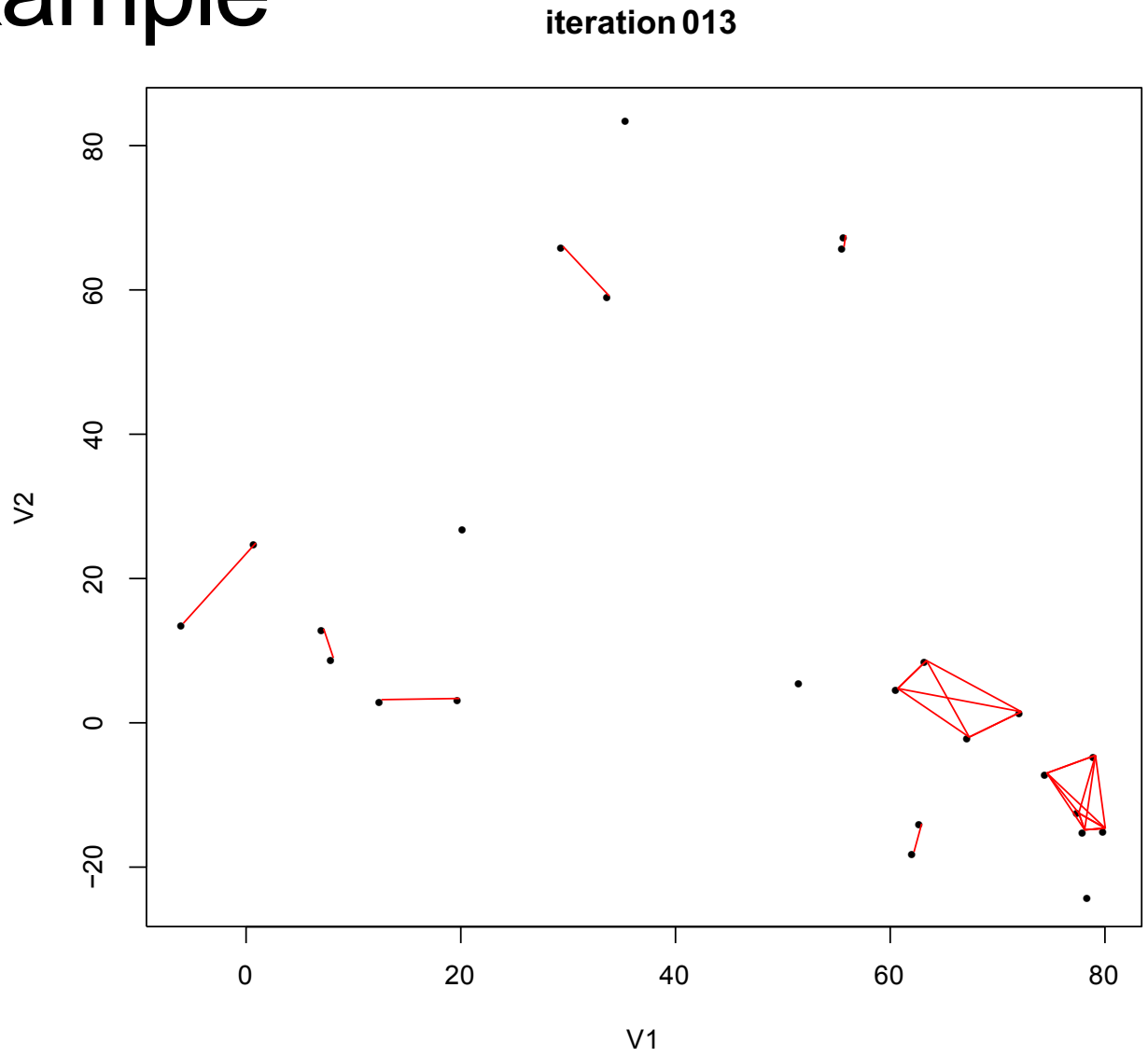
# Example



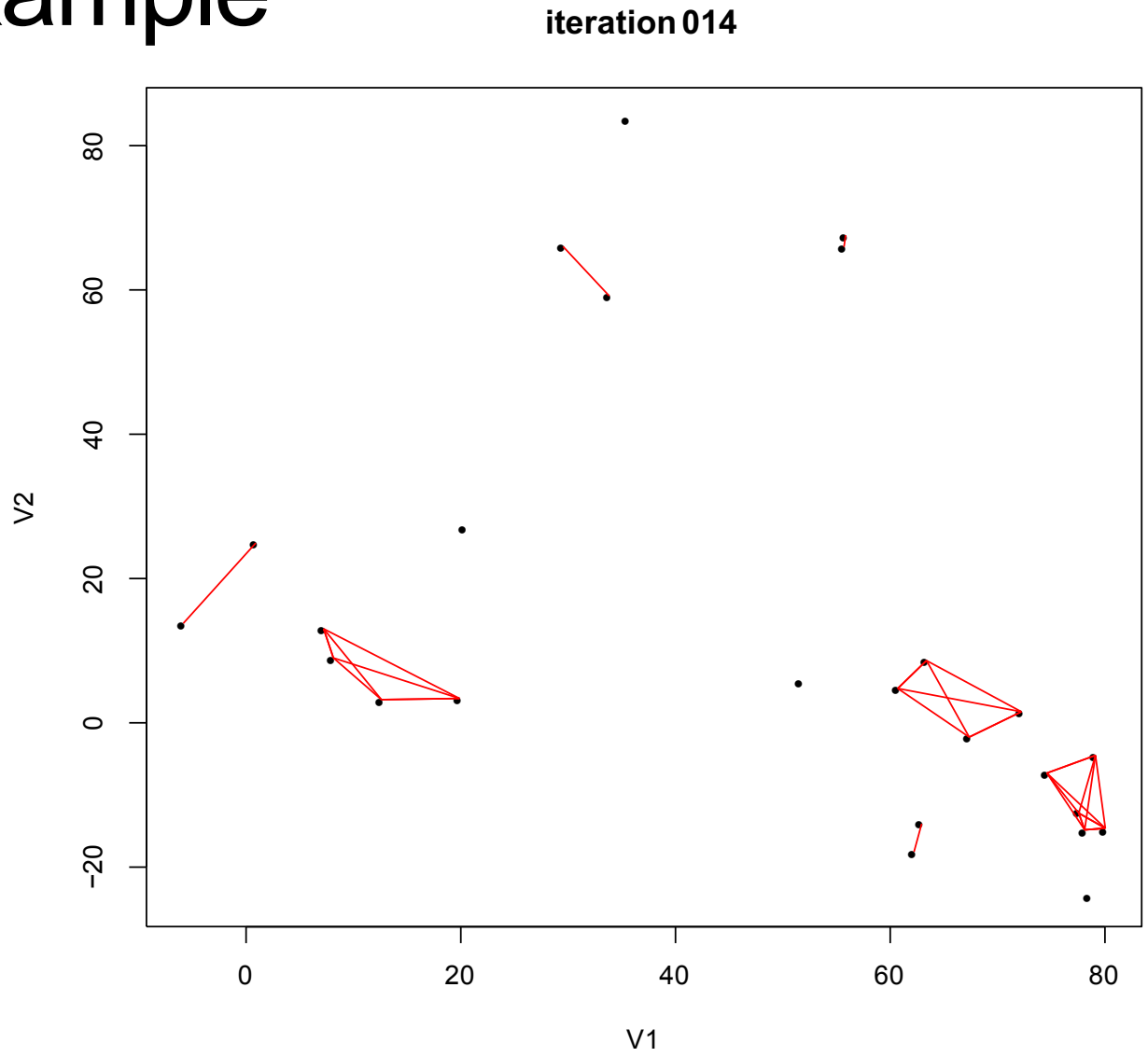
# Example



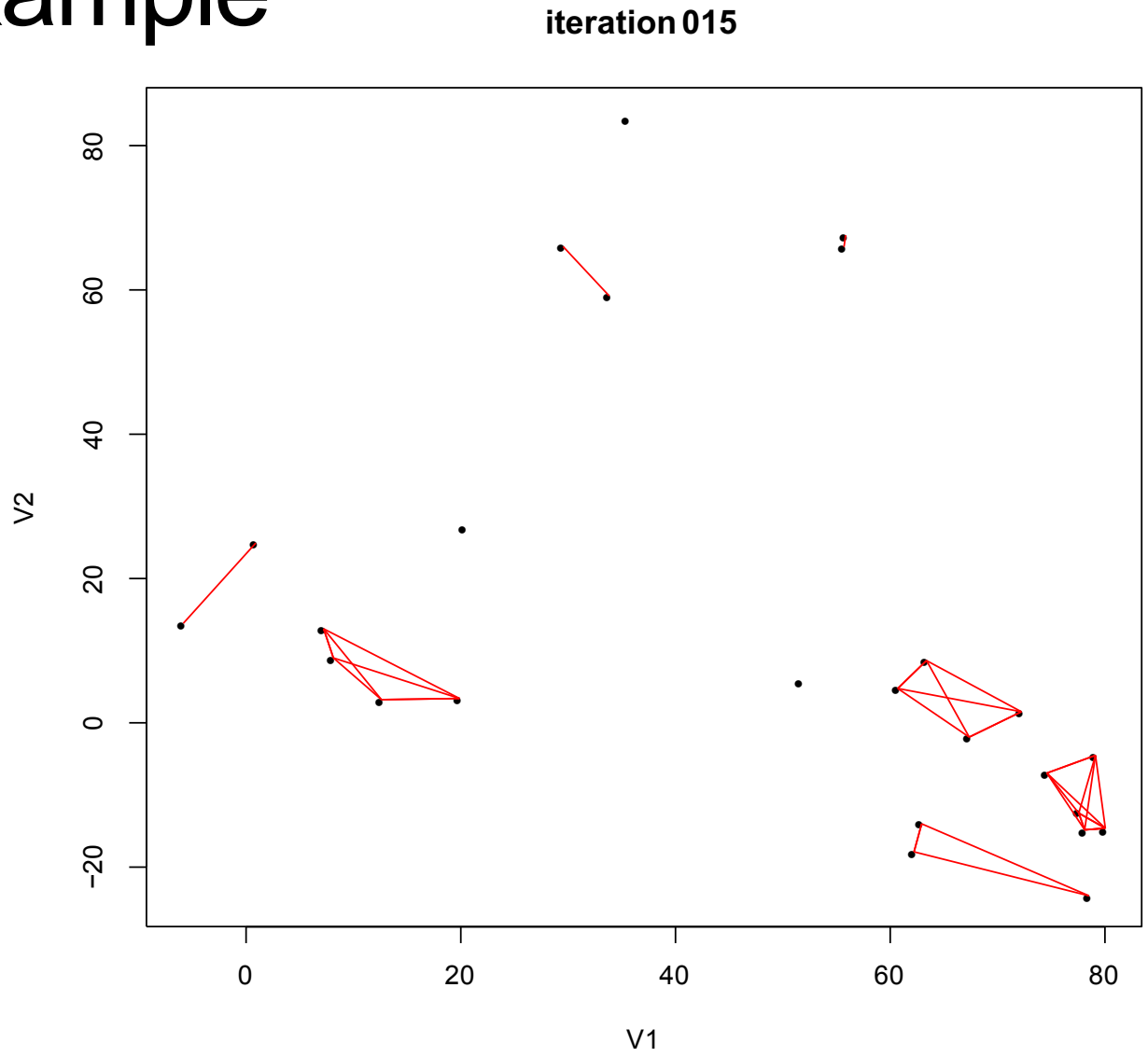
# Example



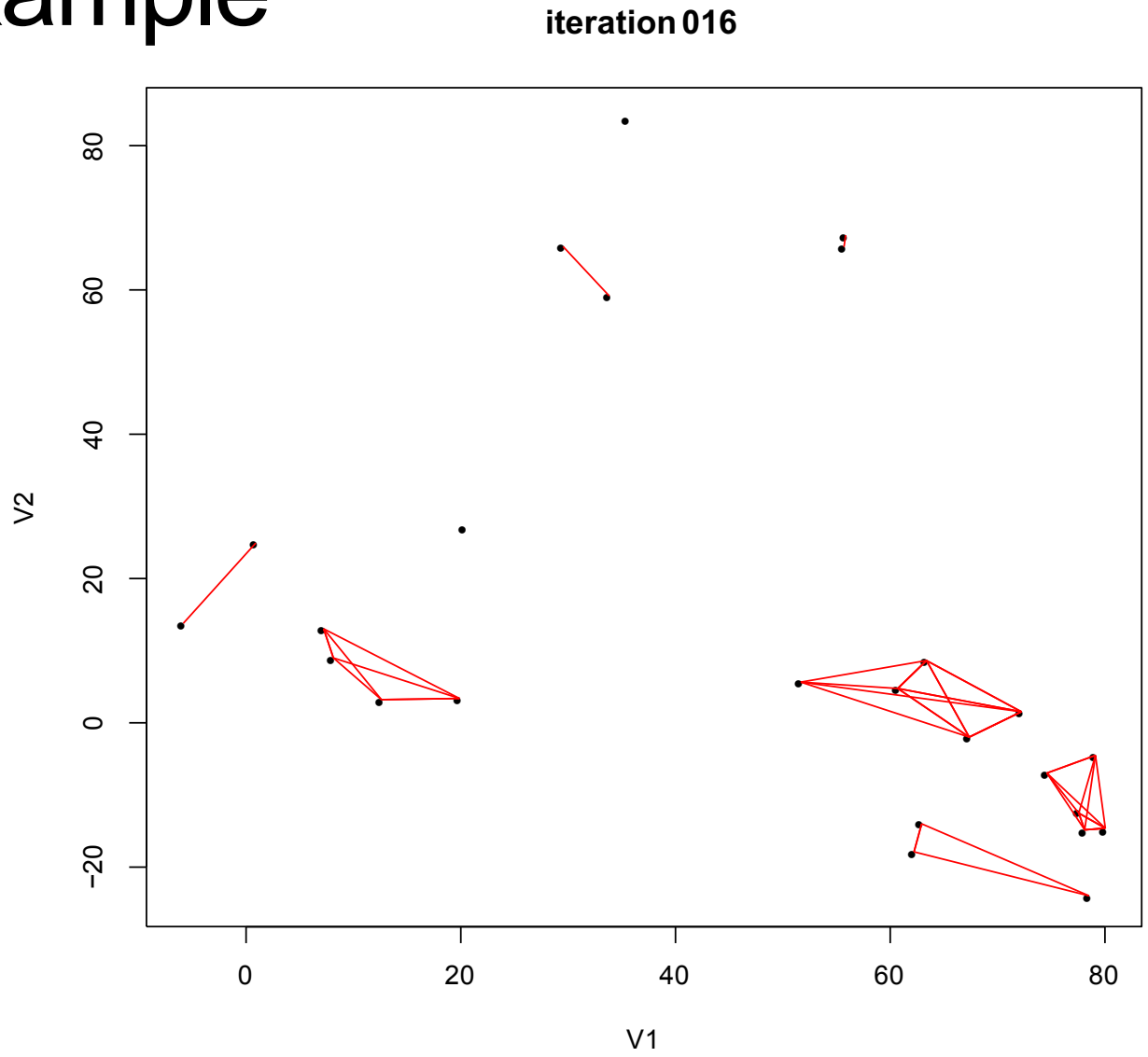
# Example



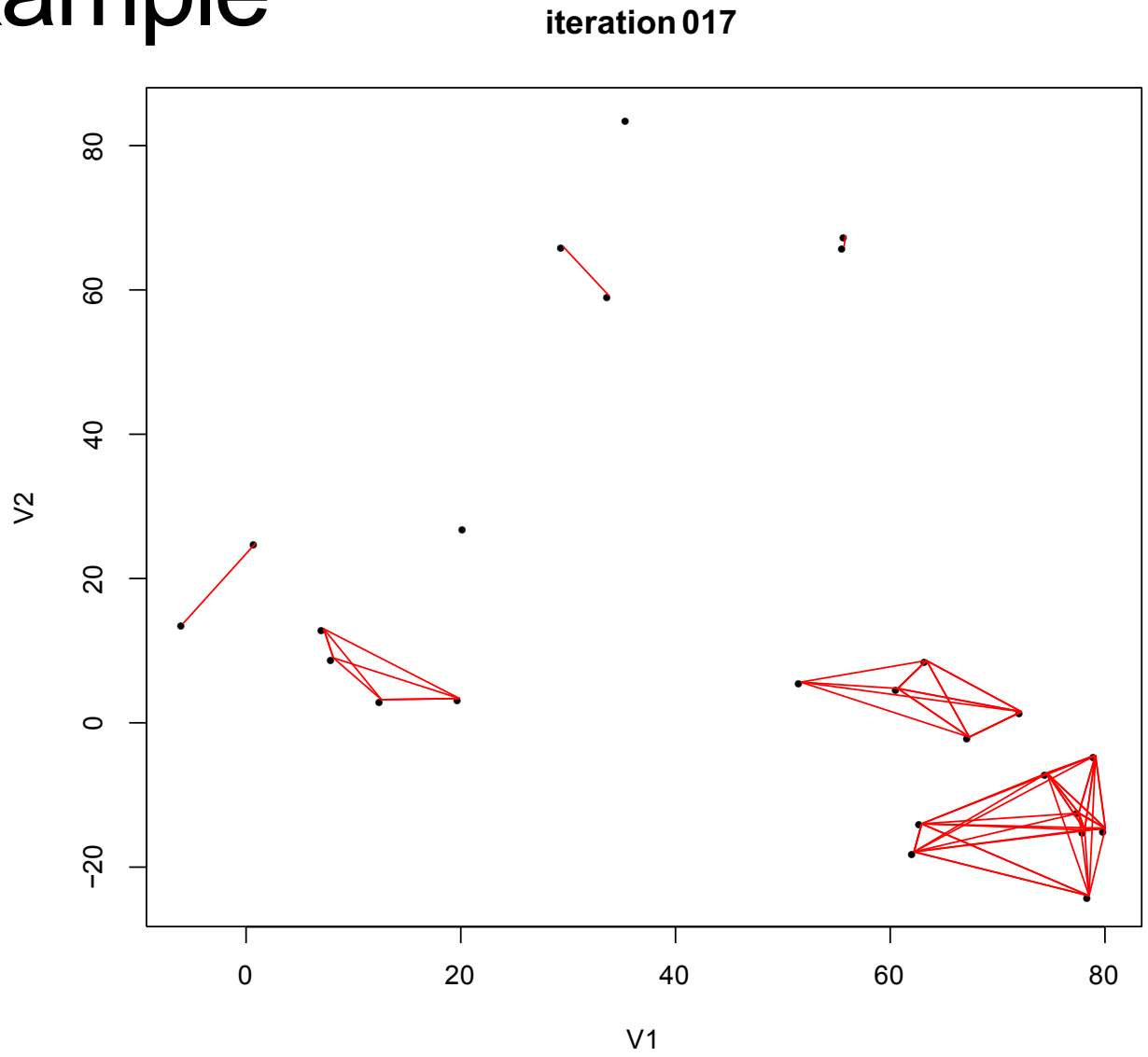
# Example



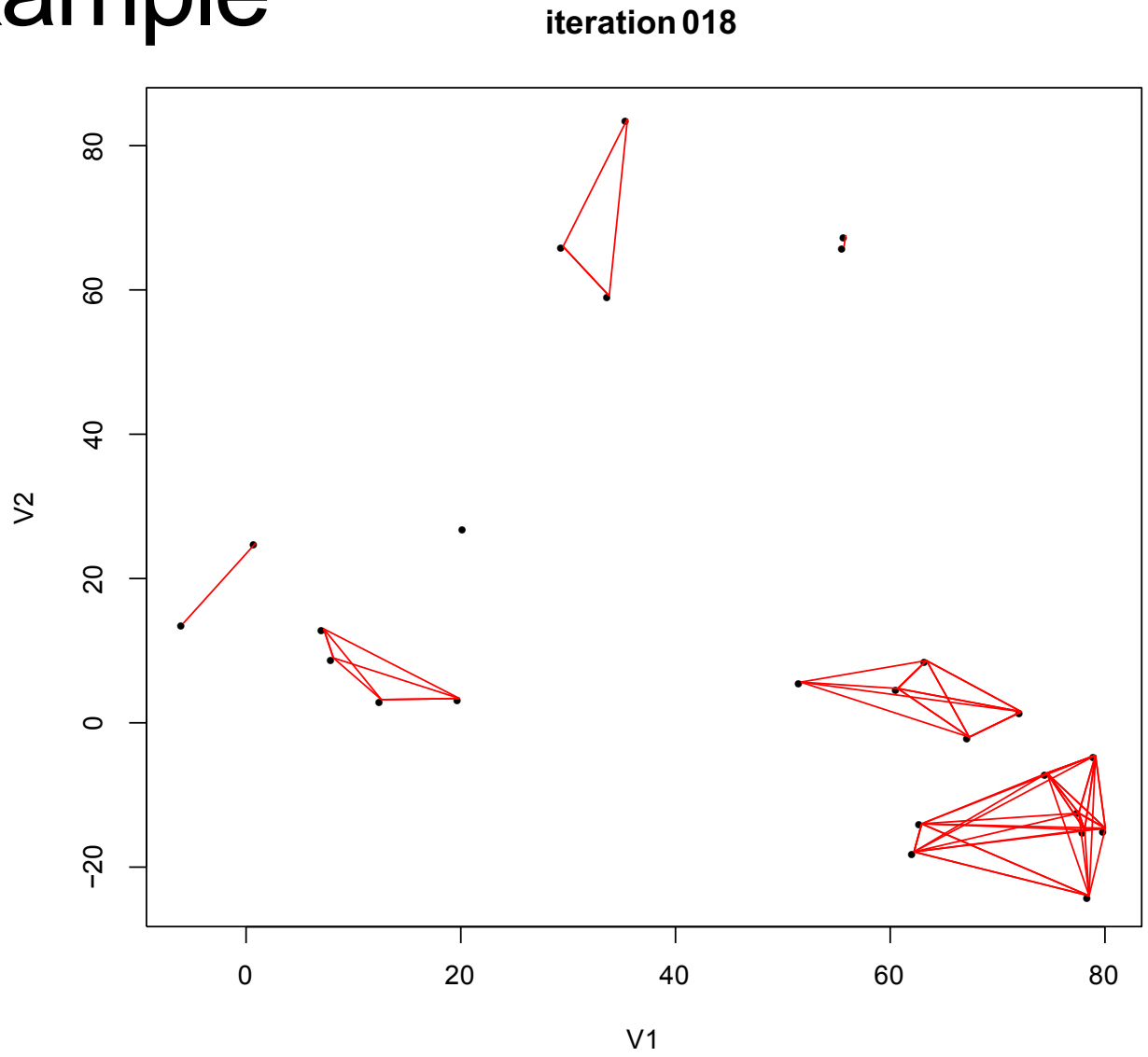
# Example



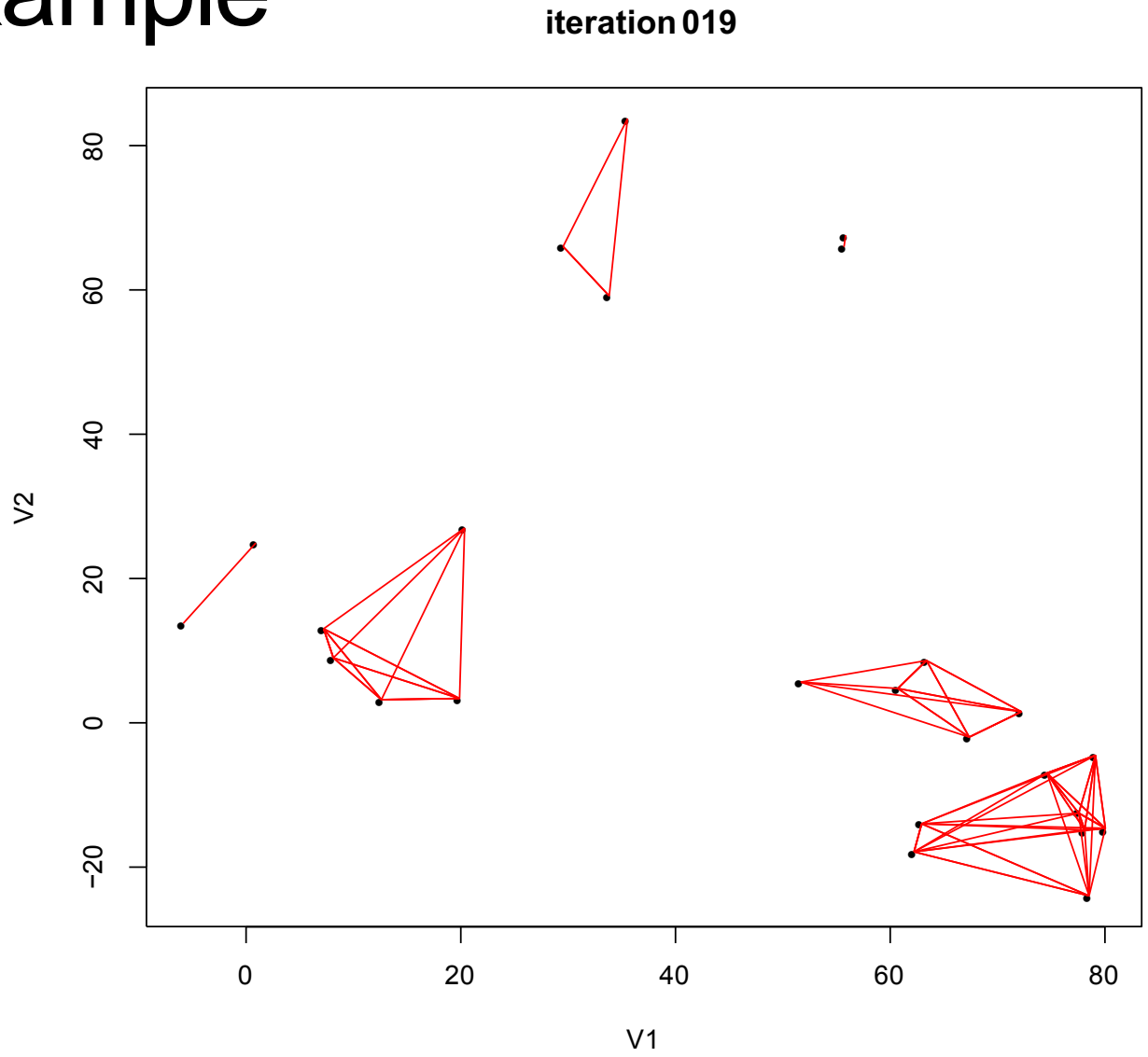
# Example



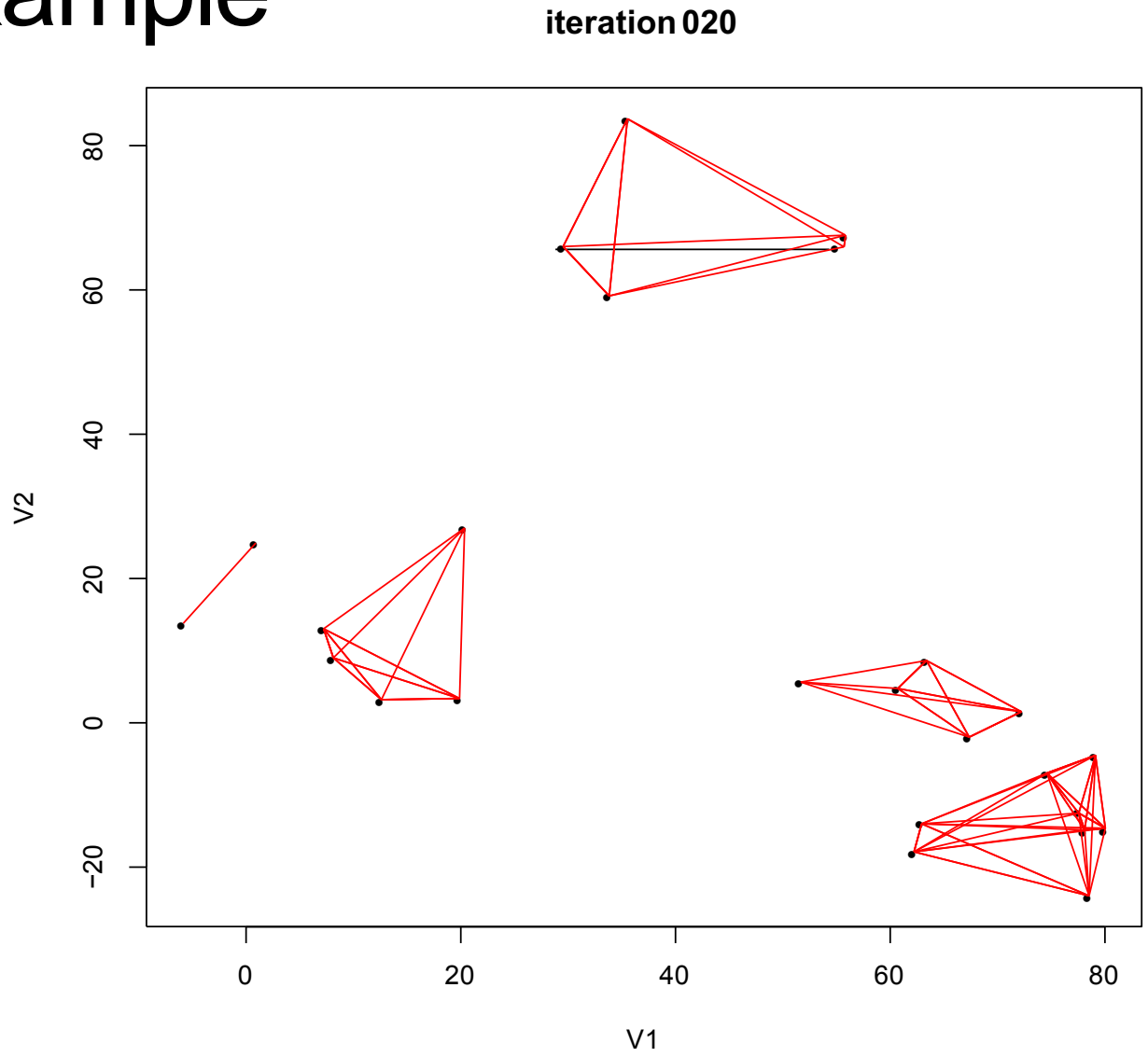
# Example



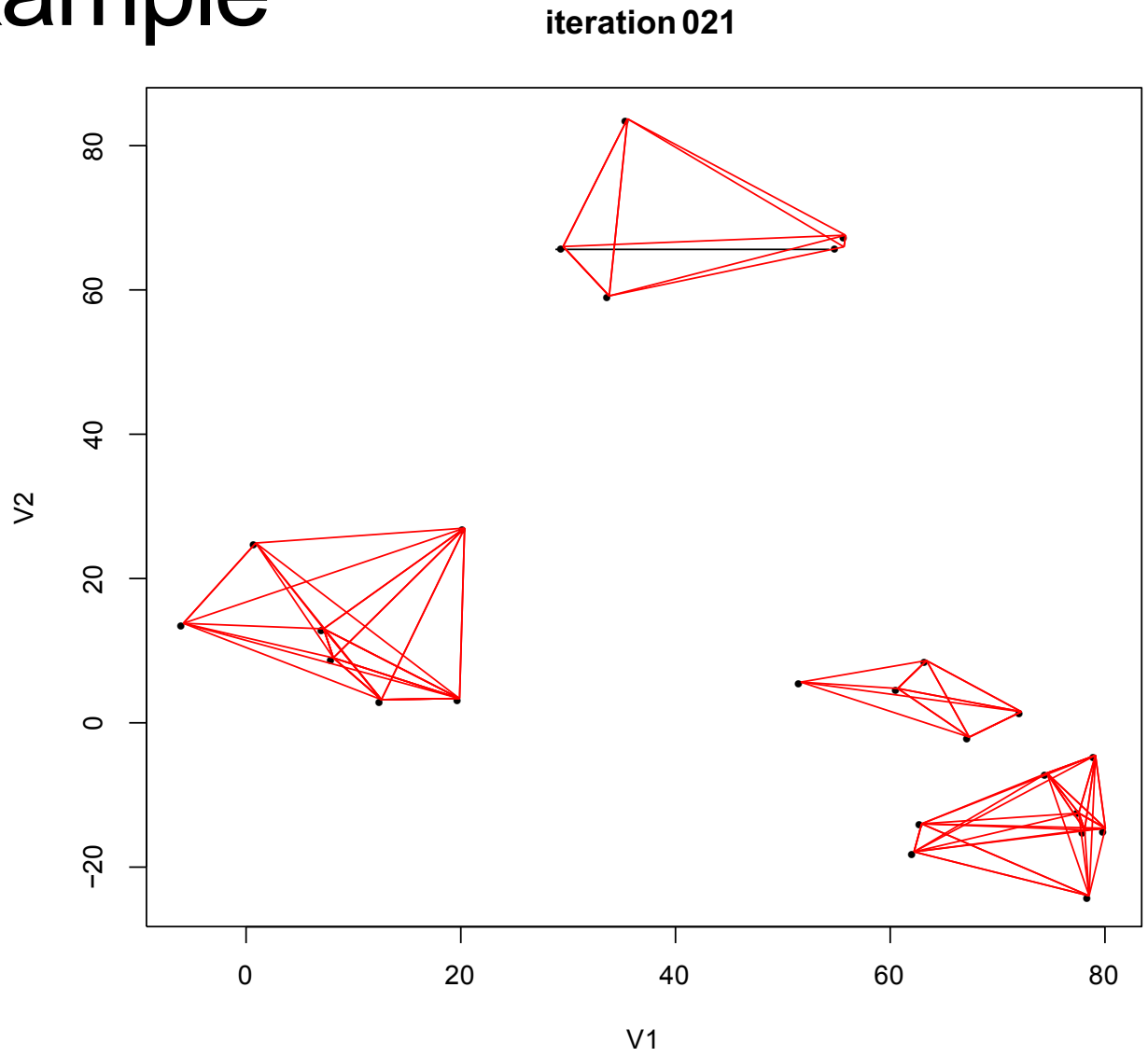
# Example



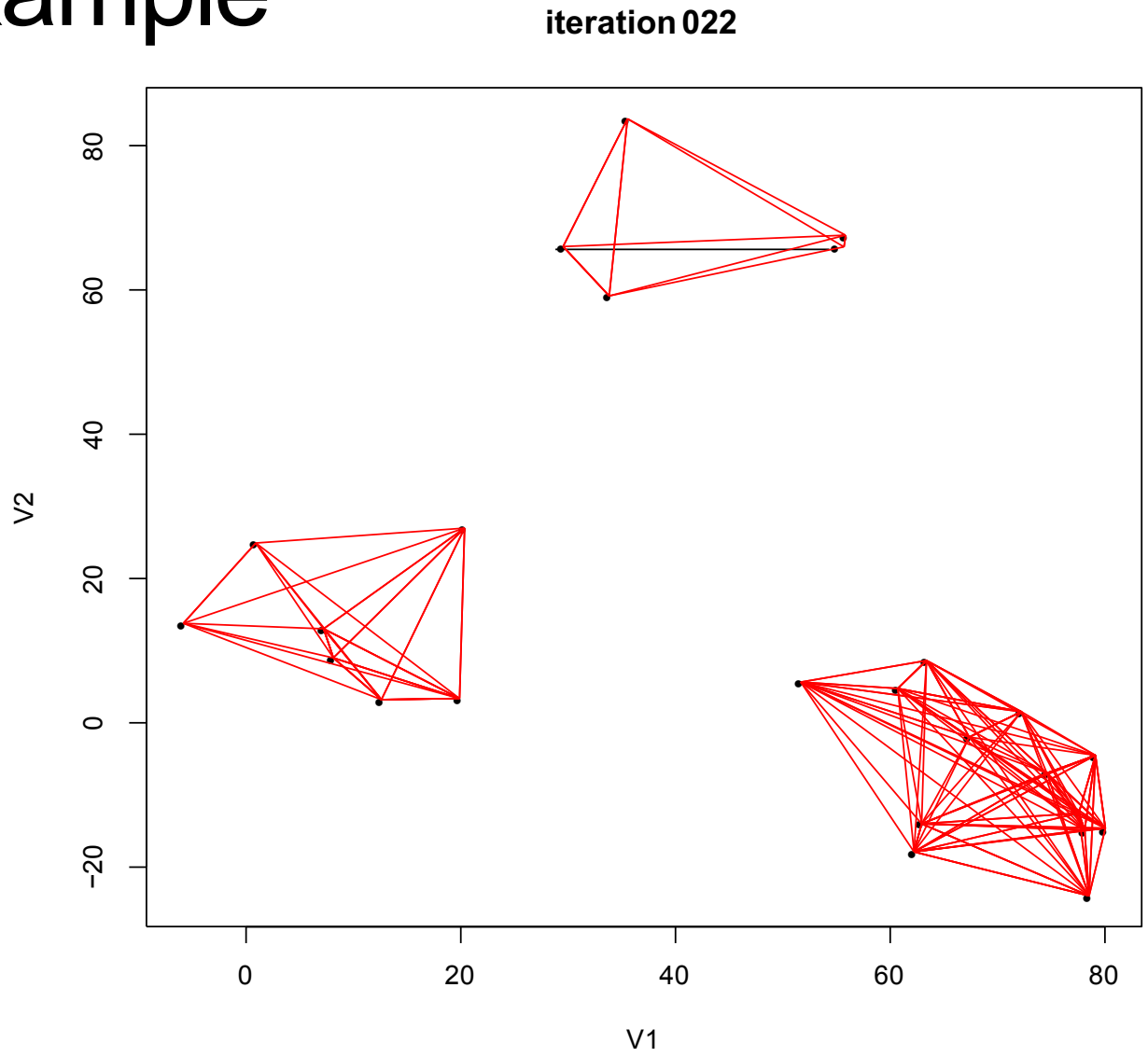
# Example



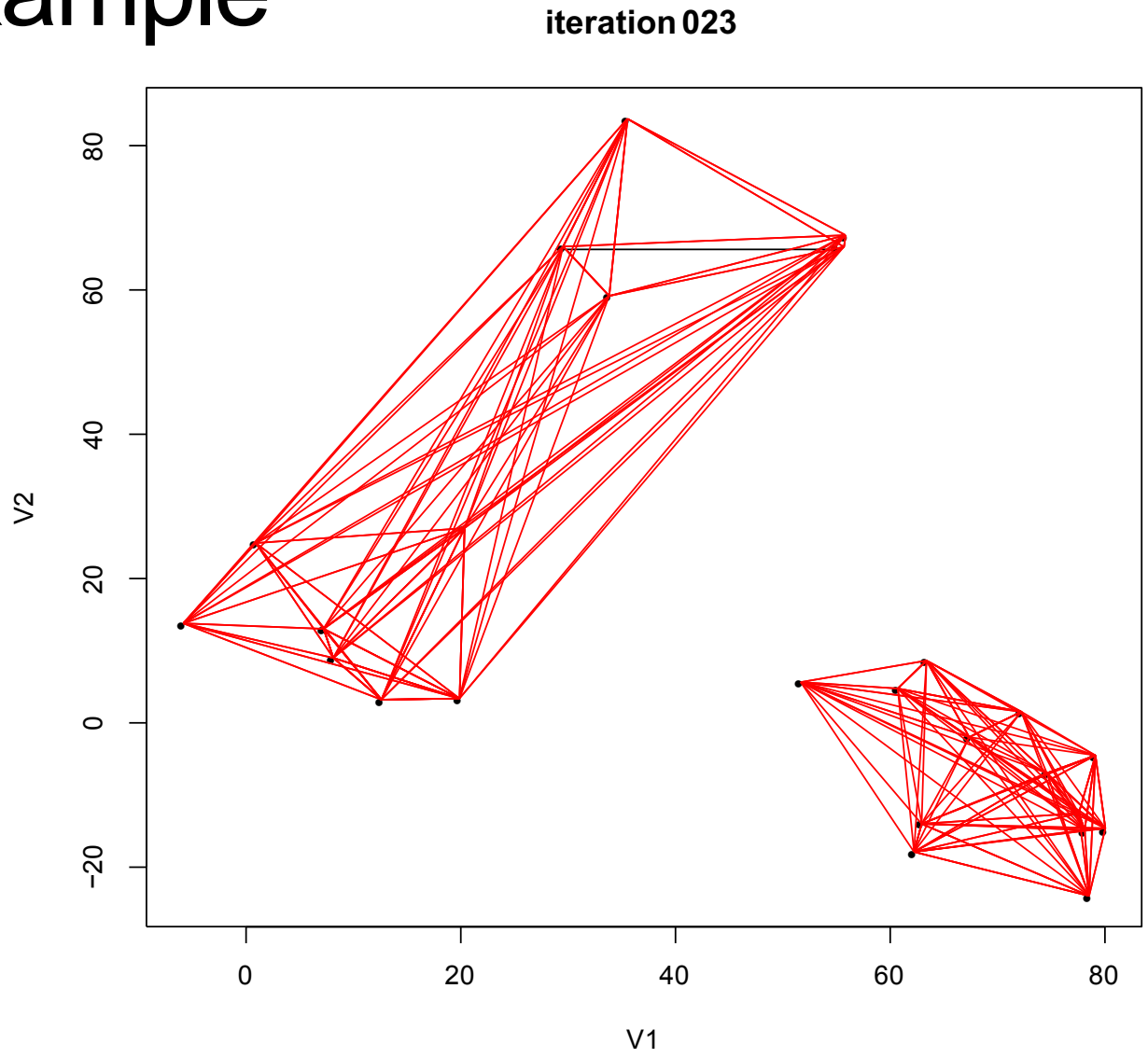
# Example



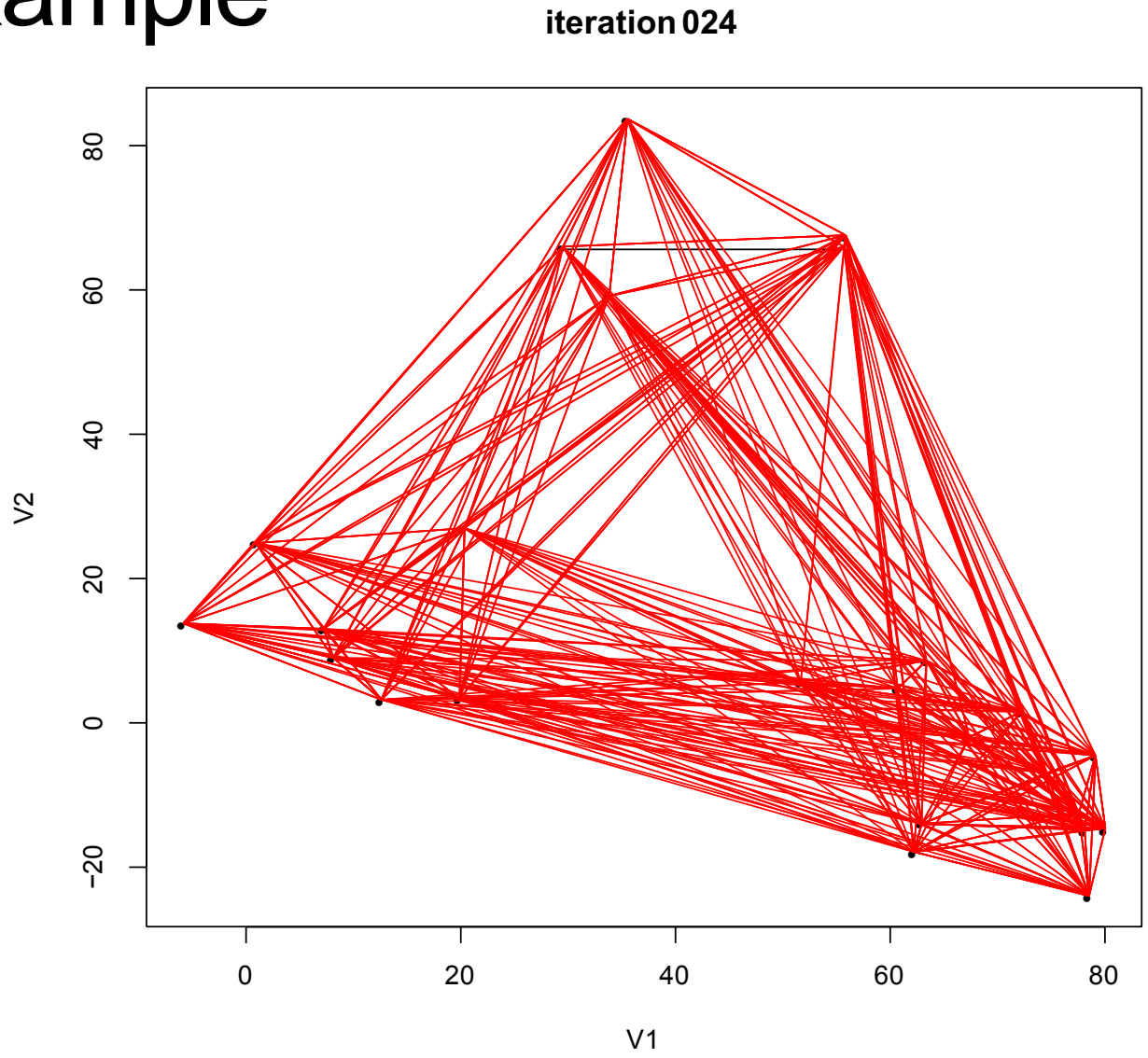
# Example



# Example

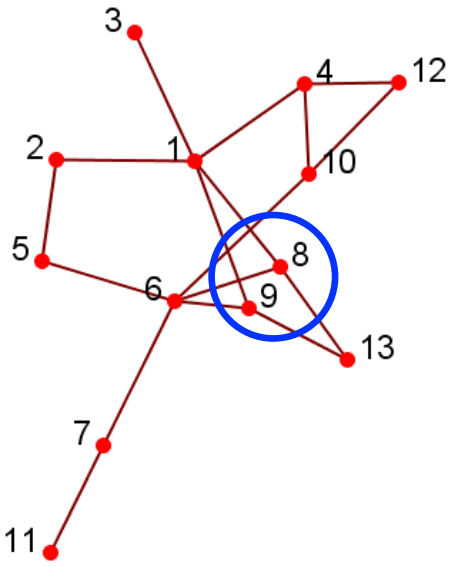


# Example



## 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

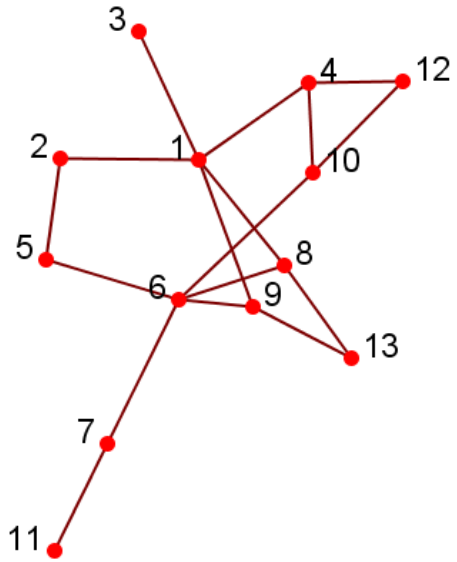


## Similarity Measures | structural equivalence or vector similarity (Cosine v Jaccard)

| | 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]

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

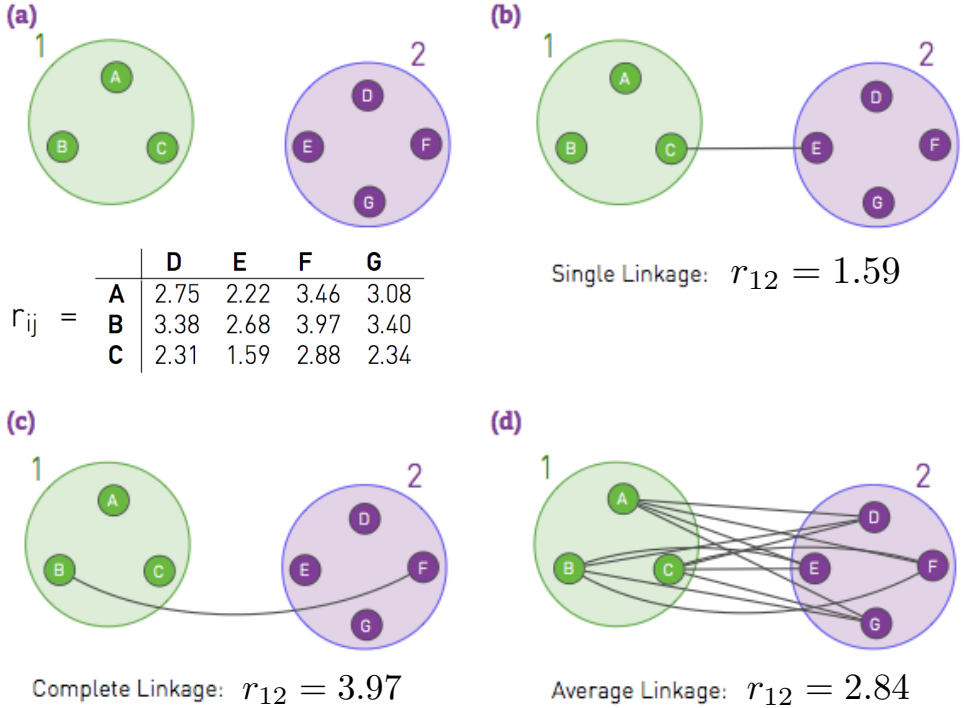
**Pearson correlation coefficient**

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

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: $s_{XY} = \frac{\sum_{x \in X, y \in Y} s_{xy}}{|X| \times |Y|}$




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

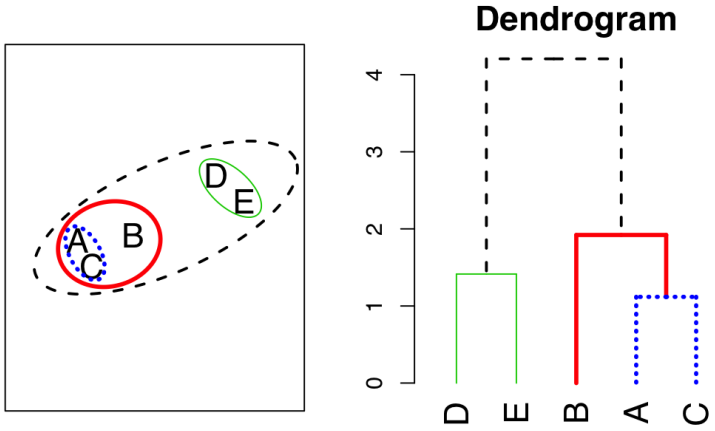
| | D | E | F | G |
|---|---|---|---|---|
| A | 2.75 | 2.22 | 3.46 | 3.08 |
| B | 3.38 | 2.68 | 3.97 | 3.40 |
| C | 2.31 | 1.59 | 2.88 | 2.34 |

**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.

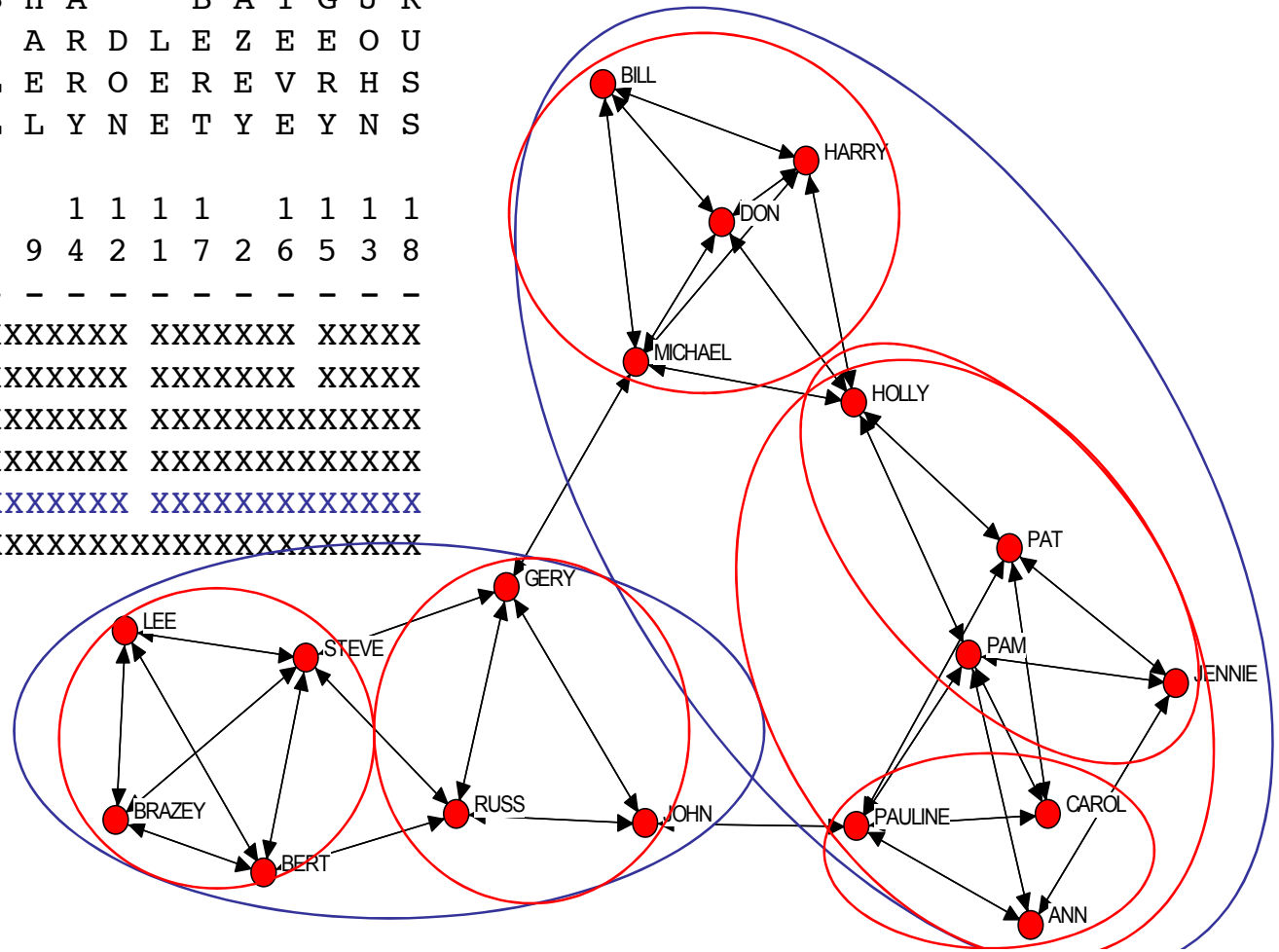
- BETTER: Compute geodesic distances first, then cluster the distance matrix (again using average method)
- Or cluster the structural equivalence matrix (tomorrow)

Geodesic Distances

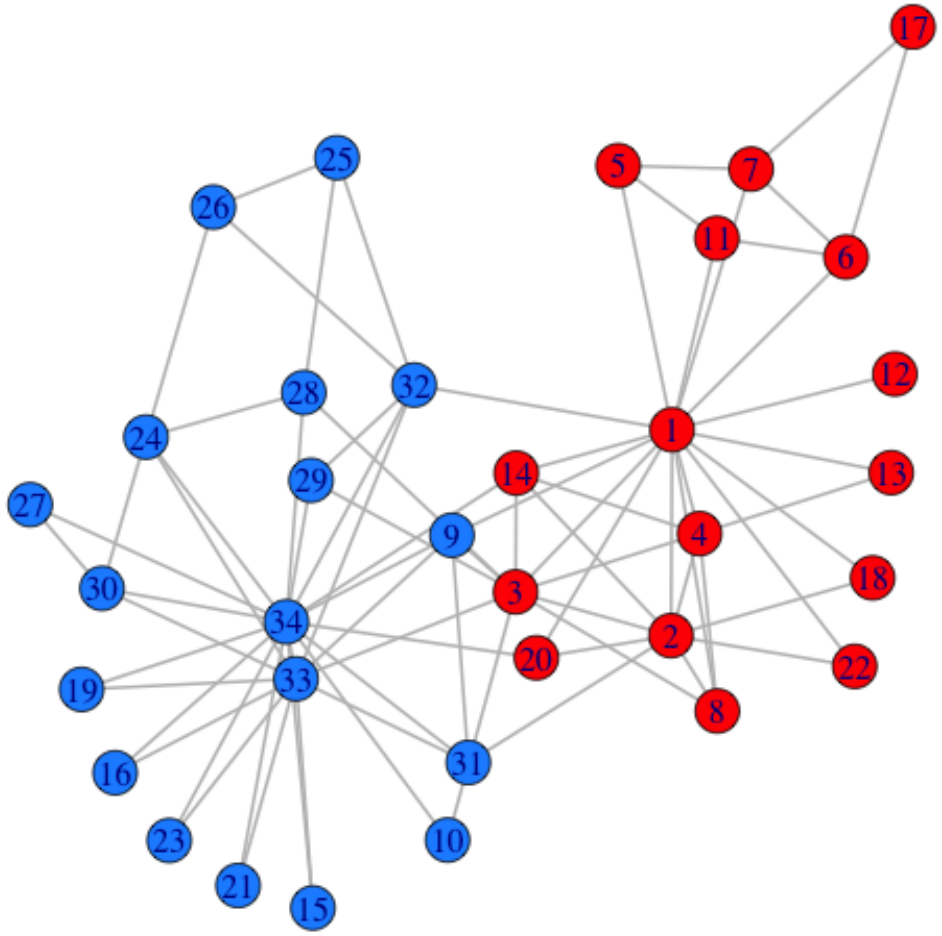
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 1 0 | 1 1 | 1 2 | 1 3 | 1 4 | 1 5 | 1 6 | 1 7 | 1 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | H | B | C | P | P | J | P | A | M | B | L | D | J | H | G | S | B | R |
| 1 | HOLLY | 0 | 4 | 2 | 1 | 1 | 2 | 2 | 2 | 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

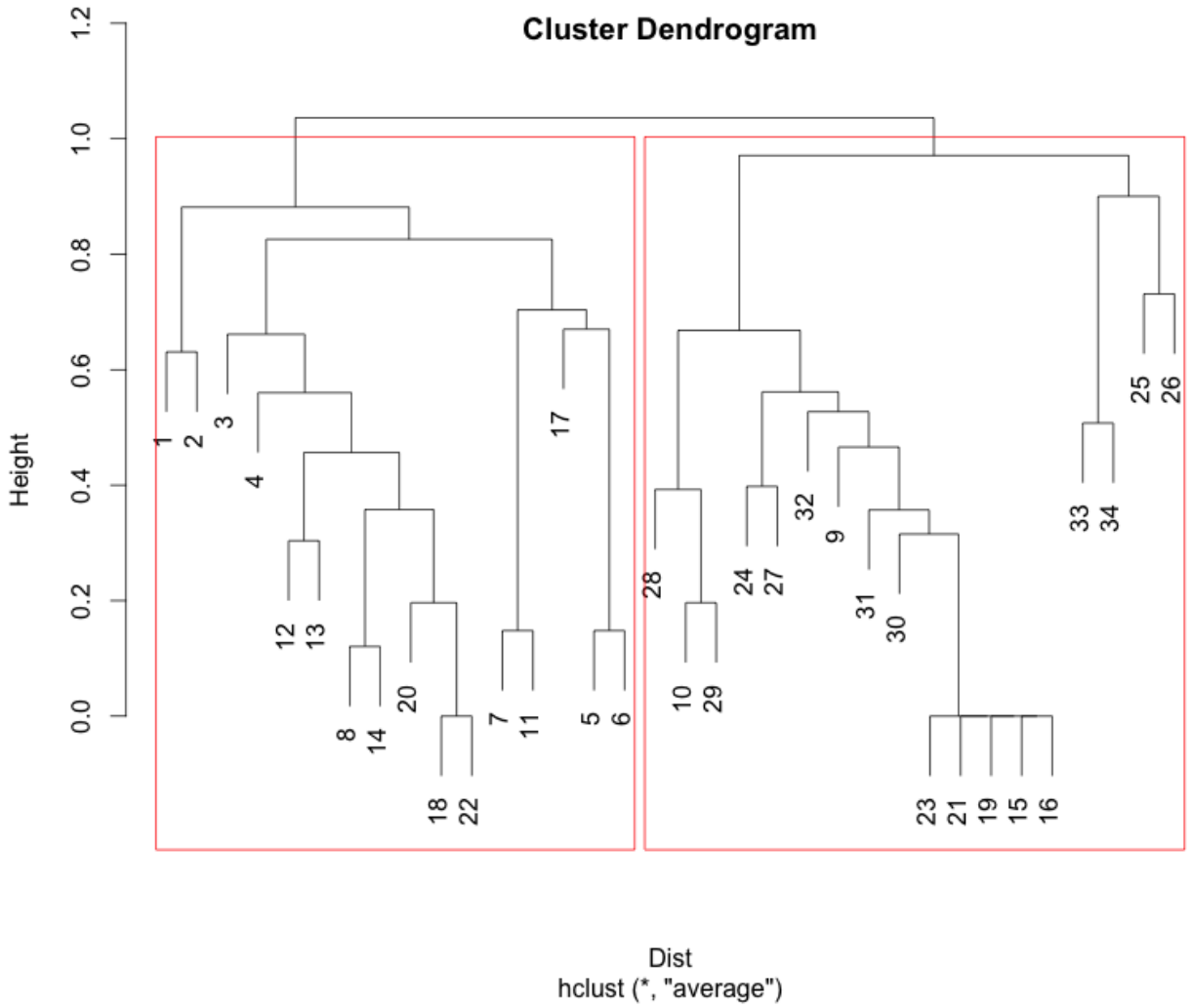
```
                P           M
                A     J     I           B
              C U   H E     C H       R S
              A L   O N   B H A     B A T G J R
          P R I P L N A I A R D L E Z E E O U
          A O N A L I N L E R O E R E V R H S
          T L E M Y E N L L Y N E T Y E Y N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



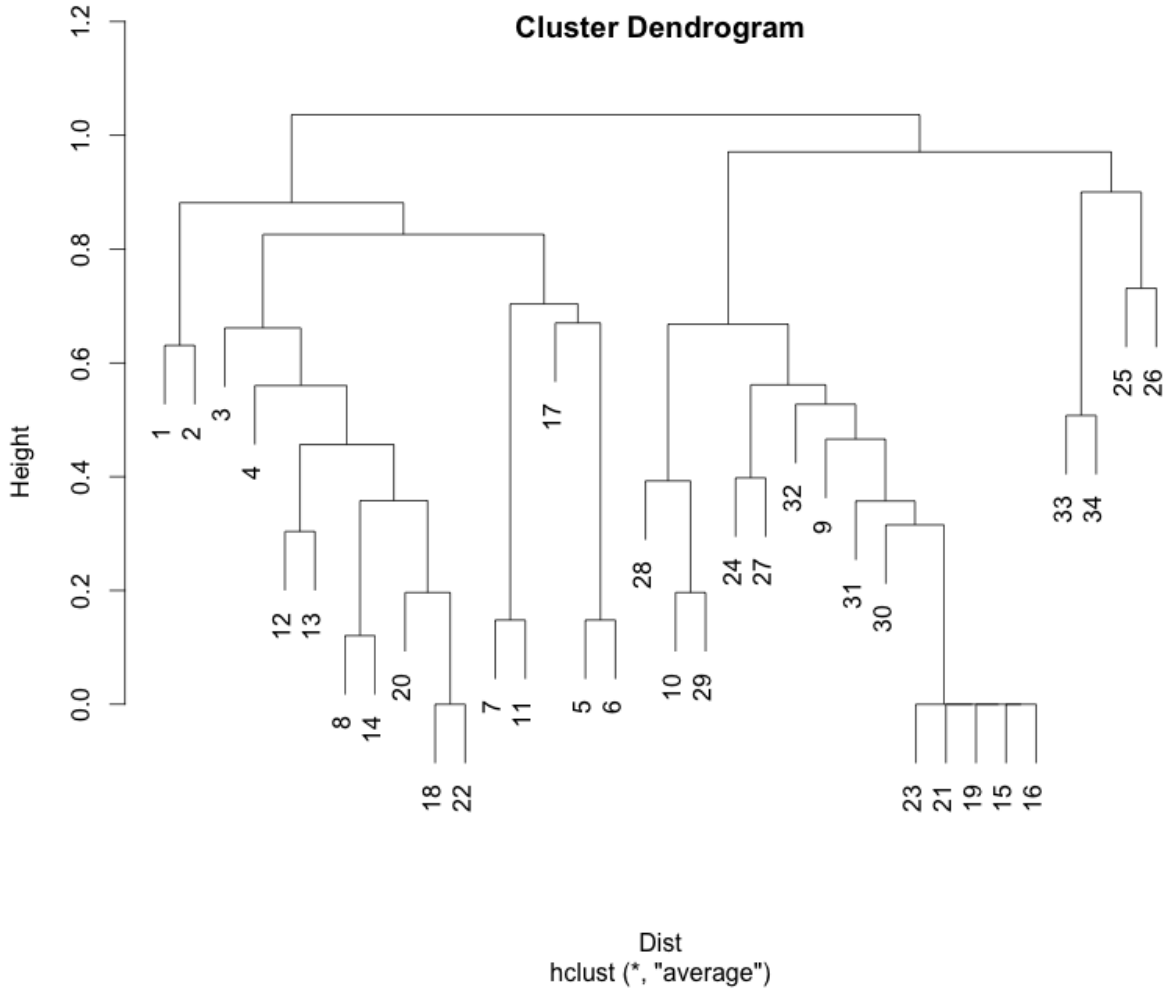
## Clustering on Node Similarities | Agglomerative Hierarchical clustering



# Clustering on Node Similarities | Agglomerative Hierarchical clustering

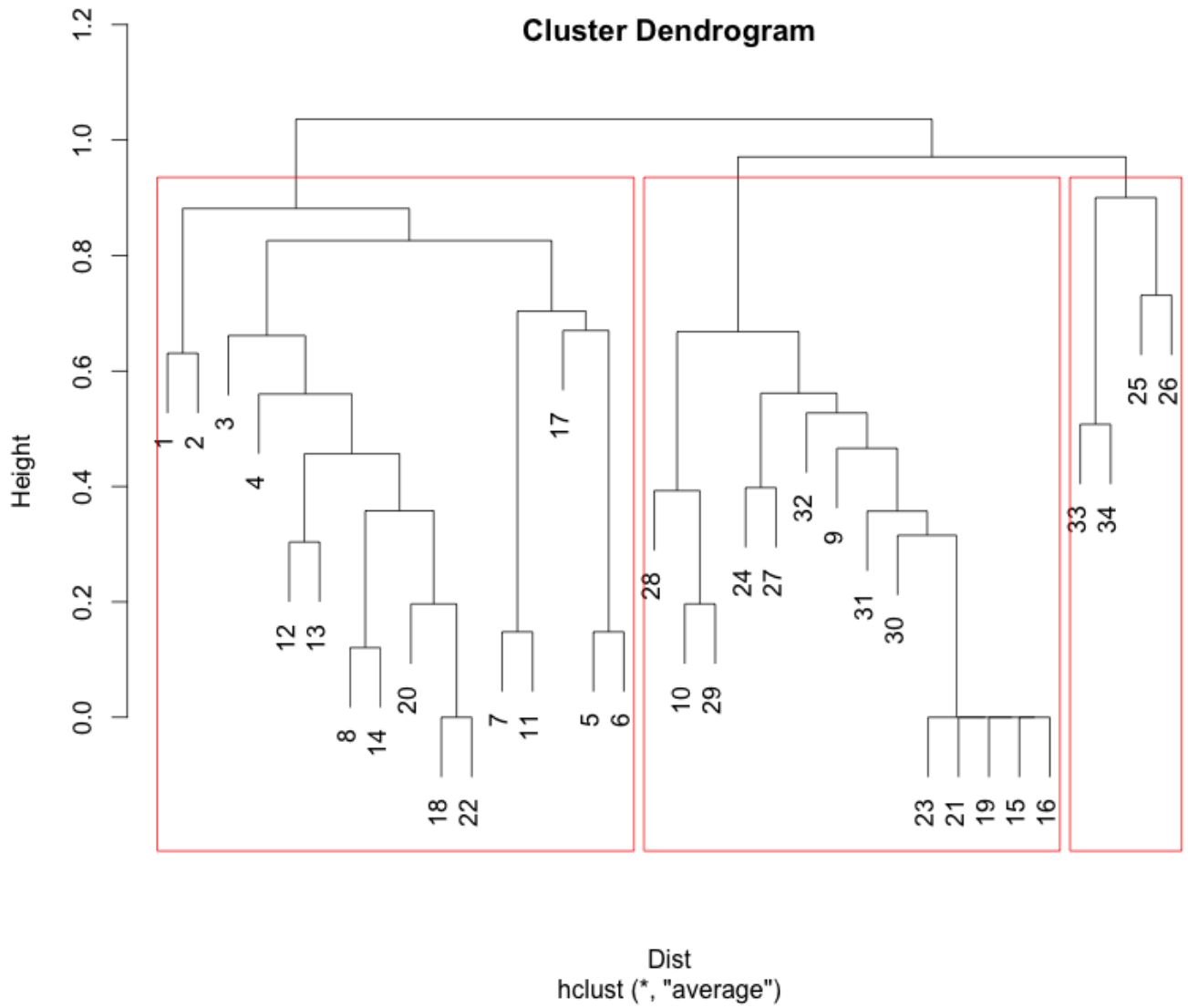


## Clustering on Node Similarities | Agglomerative Hierarchical clustering

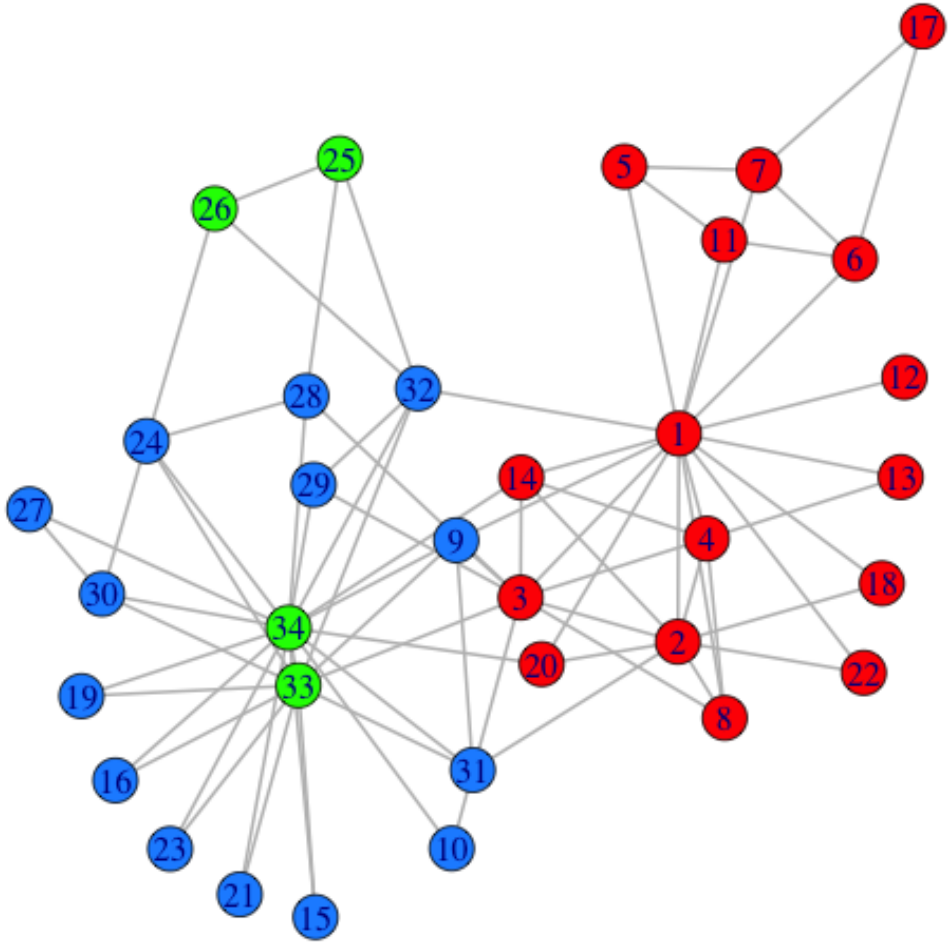



We can decide at what level we want to cut. Do we want very *fine* or very *coarse* communities?

# Clustering on Node Similarities | Agglomerative Hierarchical clustering



## Clustering on Node Similarities | Agglomerative Hierarchical clustering

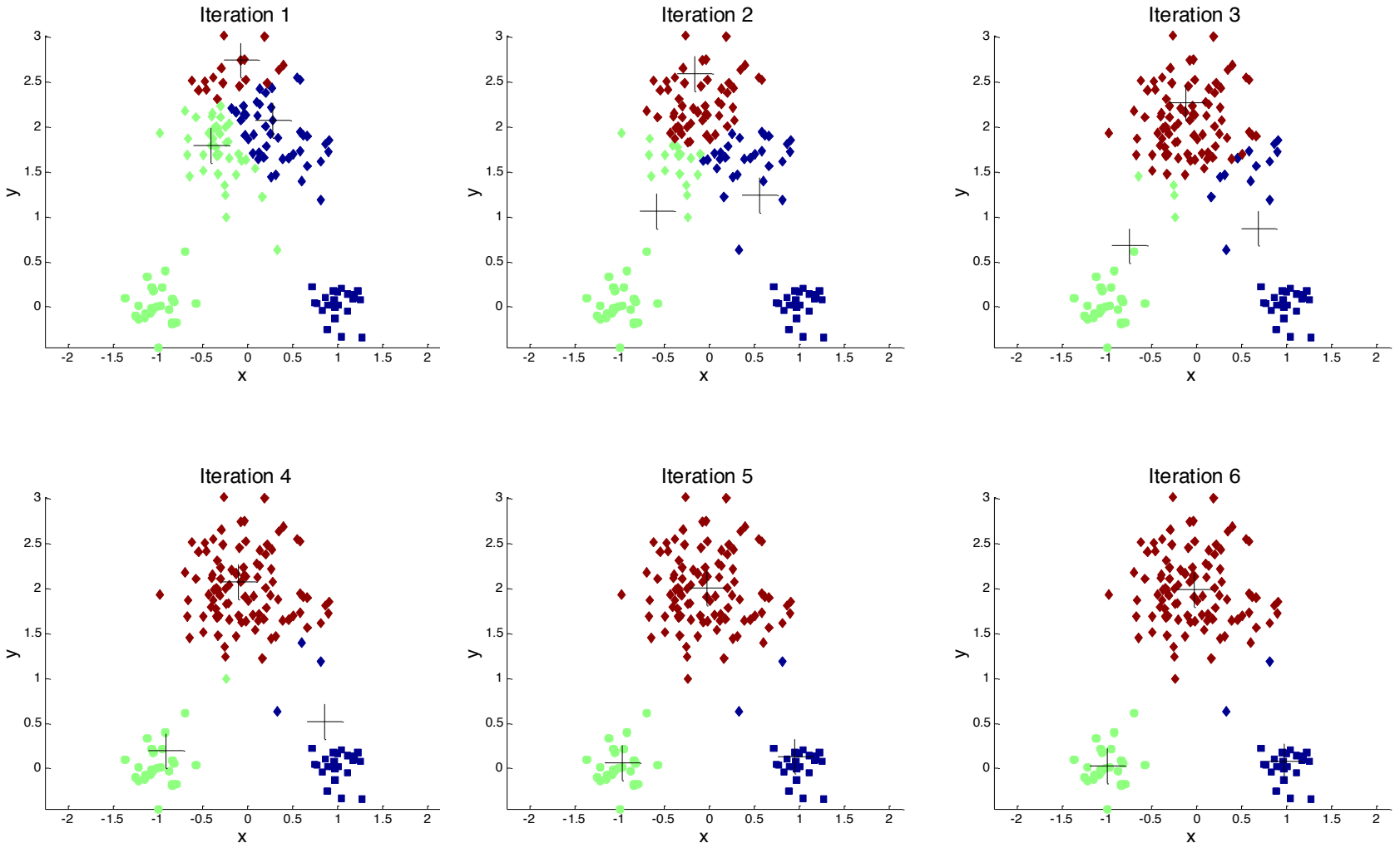


- K-means Clustering Algorithm
  - Each cluster is associated with a centroid (center point)
  - Each node is assigned to the cluster with the closest centroid

**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



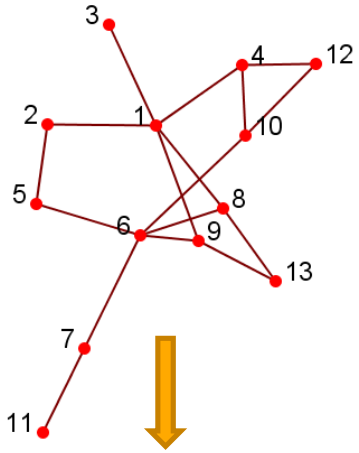
## Node Similarity| Multidimensional Scaling

- **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 $\Delta$), V the top-k eigenvectors
  - Solution: $S = V\Lambda^{1/2}$
- Apply k-means to S to obtain clusters

# Node Similarity| Multidimensional Scaling



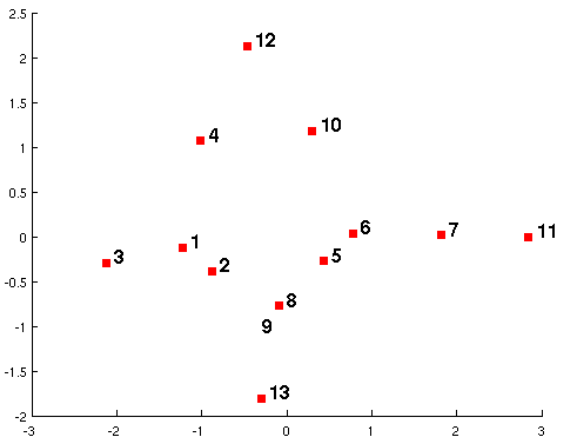

## Geodesic Distance Matrix

| | 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 |

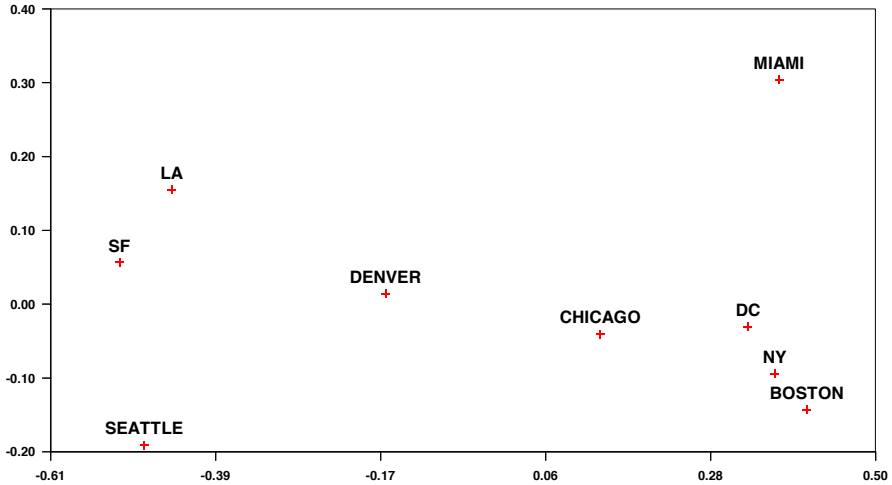


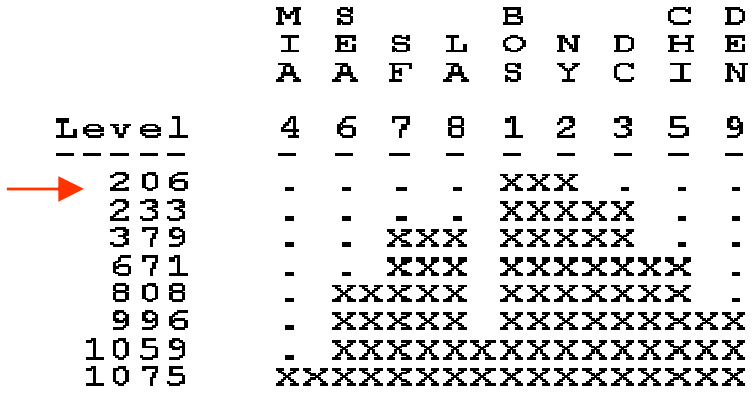

k-means

1, 2, 3, 4, 10, 12

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

# Node Similarity| Multidimensional Scaling

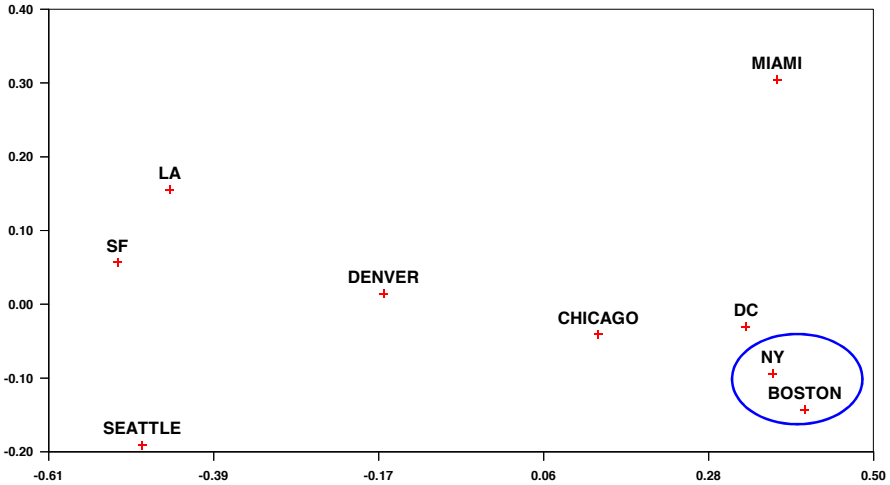


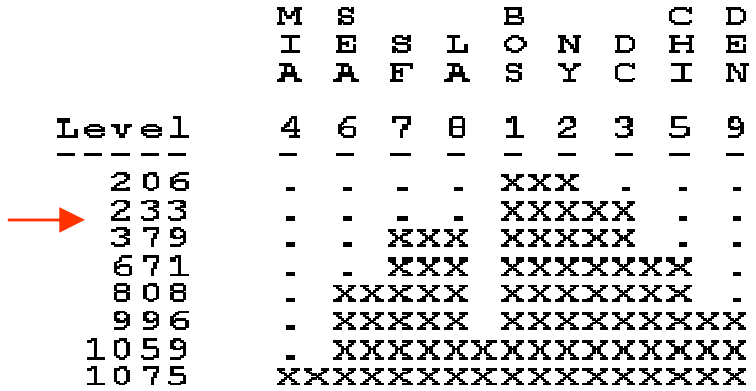




| | BOS | NY | DC | MIA | CHI | SEA | SF | LA | DEN |
|---|---|---|---|---|---|---|---|---|---|
| BOS | 0 | 206 | 429 | 1504 | 963 | 2976 | 3095 | 2979 | 1949 |
| NY | 206 | 0 | 233 | 1308 | 802 | 2815 | 2934 | 2786 | 1771 |
| DC | 429 | 233 | 0 | 1075 | 671 | 2684 | 2799 | 2631 | 1616 |
| MIA | 1504 | 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

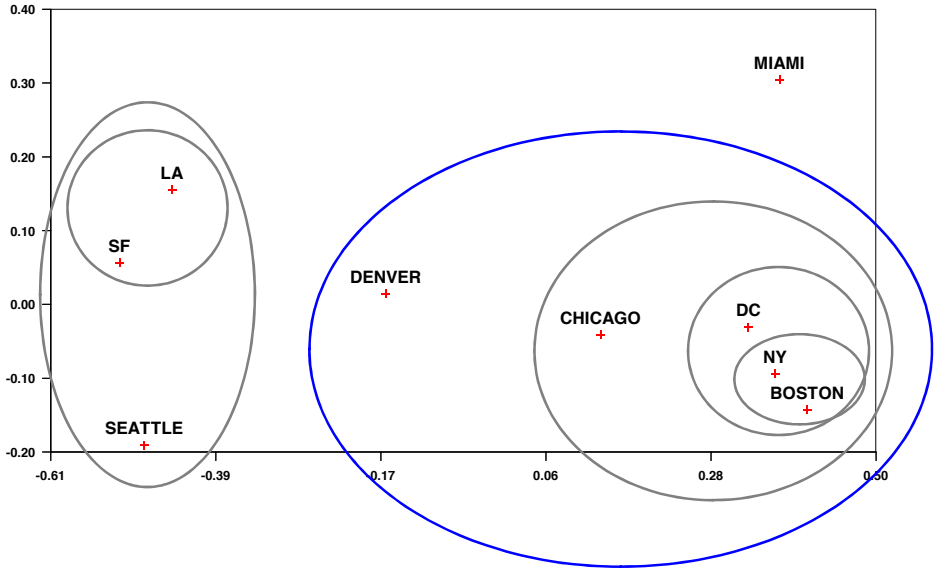


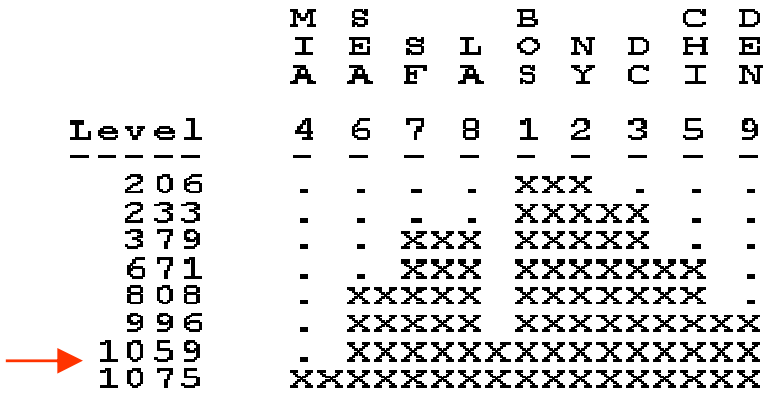




| | BOS NY | DC | MIA | CHI | SEA | SF | LA | DEN |
|---|---|---|---|---|---|---|---|---|
| BOS/ NY | 0 | 233 | 1308 | 802 | 2815 | 2934 | 2786 | 1771 |
| DC | 233 | 0 | 1075 | 671 | 2684 | 2799 | 2631 | 1616 |
| MIA | 1308 | 1075 | 0 | 1329 | 3273 | 3053 | 2687 | 2037 |
| CHI | 802 | 671 | 1329 | 0 | 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.

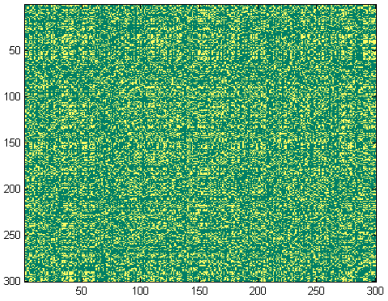
# Node Similarity| Multidimensional Scaling



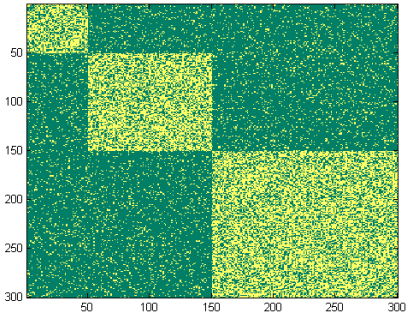




| | BOS/ NY/D C/CHI /DEN | MIA | SF/LA /SEA |
|---|---|---|---|
| BOS/NY/DC/ CHI/DEN | 0 | 1075 | 1059 |
| MIA | 1075 | 0 | 2687 |
| SF/LA/SEA | 1059 | 2687 | 0 |

## 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)

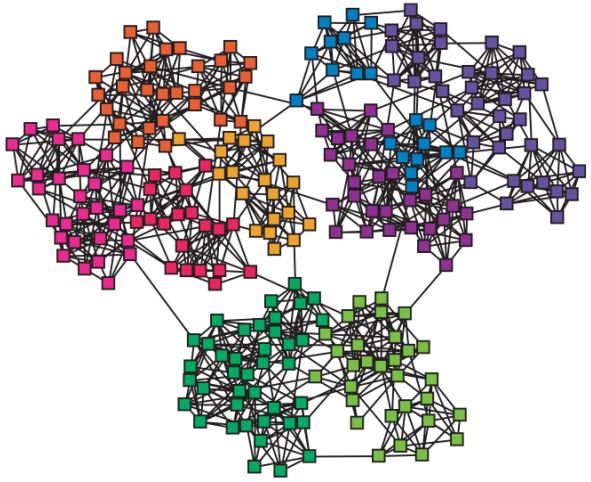
$$B_n = \sum_{k=1}^{n} S(n,k)$$

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

# 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 $C_B(e)$;

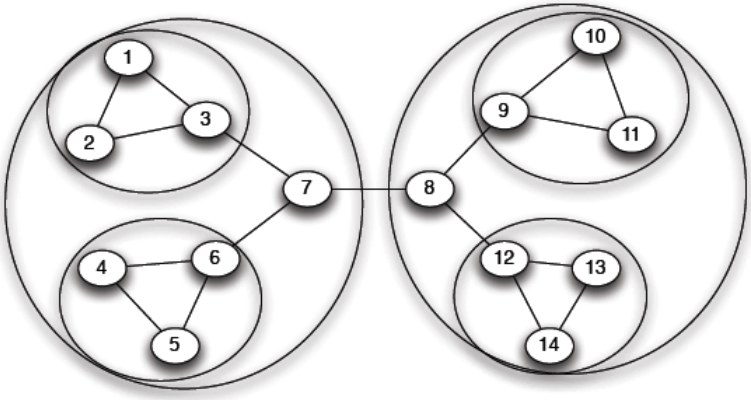
  remove edge $e_i$ with largest $C_B(e_i)$ ;

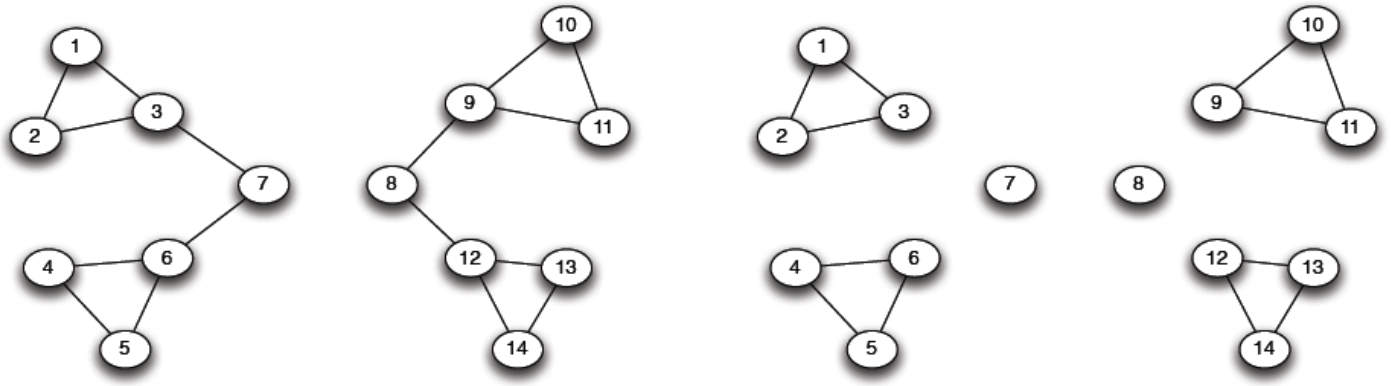
**until** *edges left*;

Construct communities by progressively removing edges

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

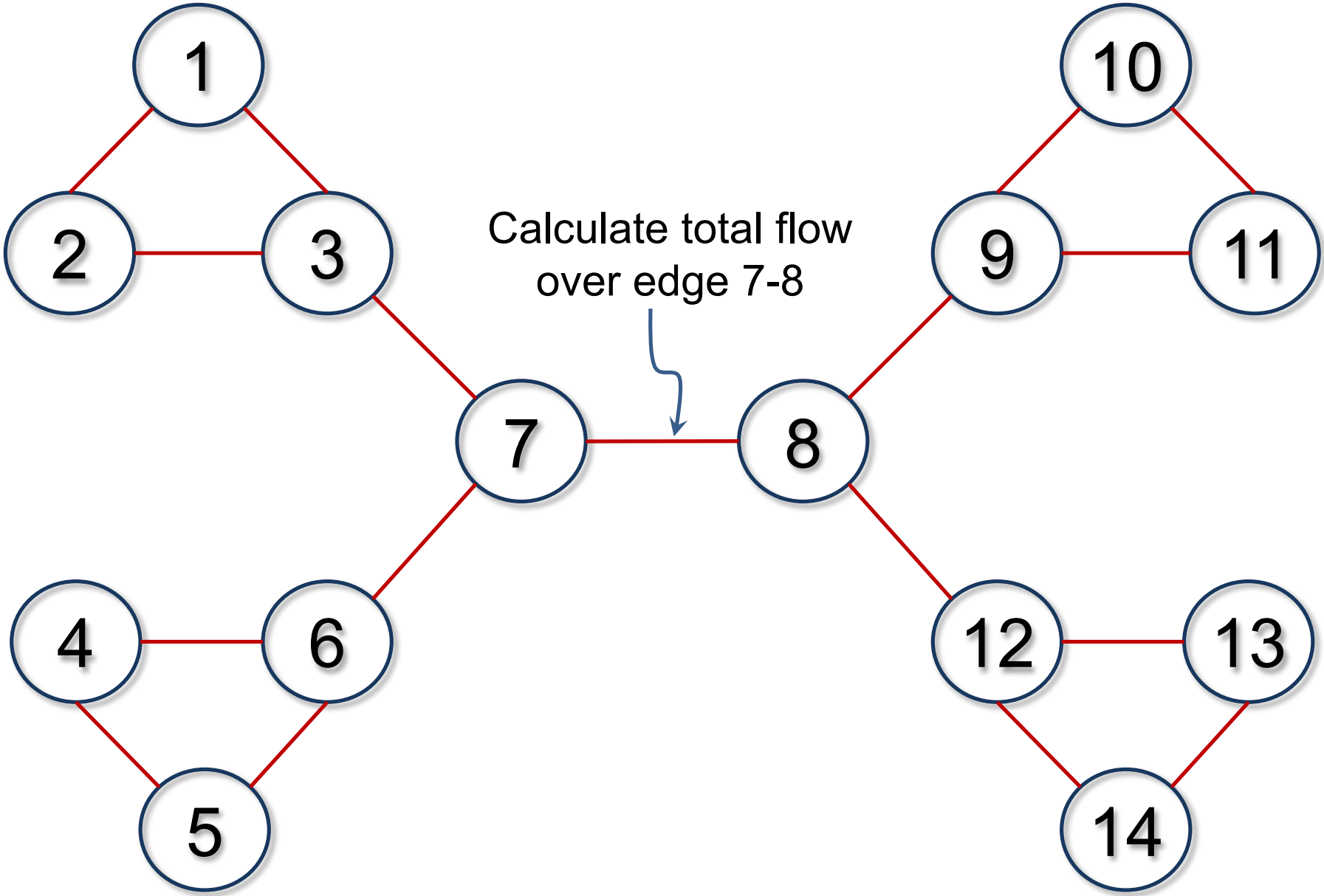
- successively remove edges of highest betweenness (the bridges, or local bridges), breaking up the network into separate components

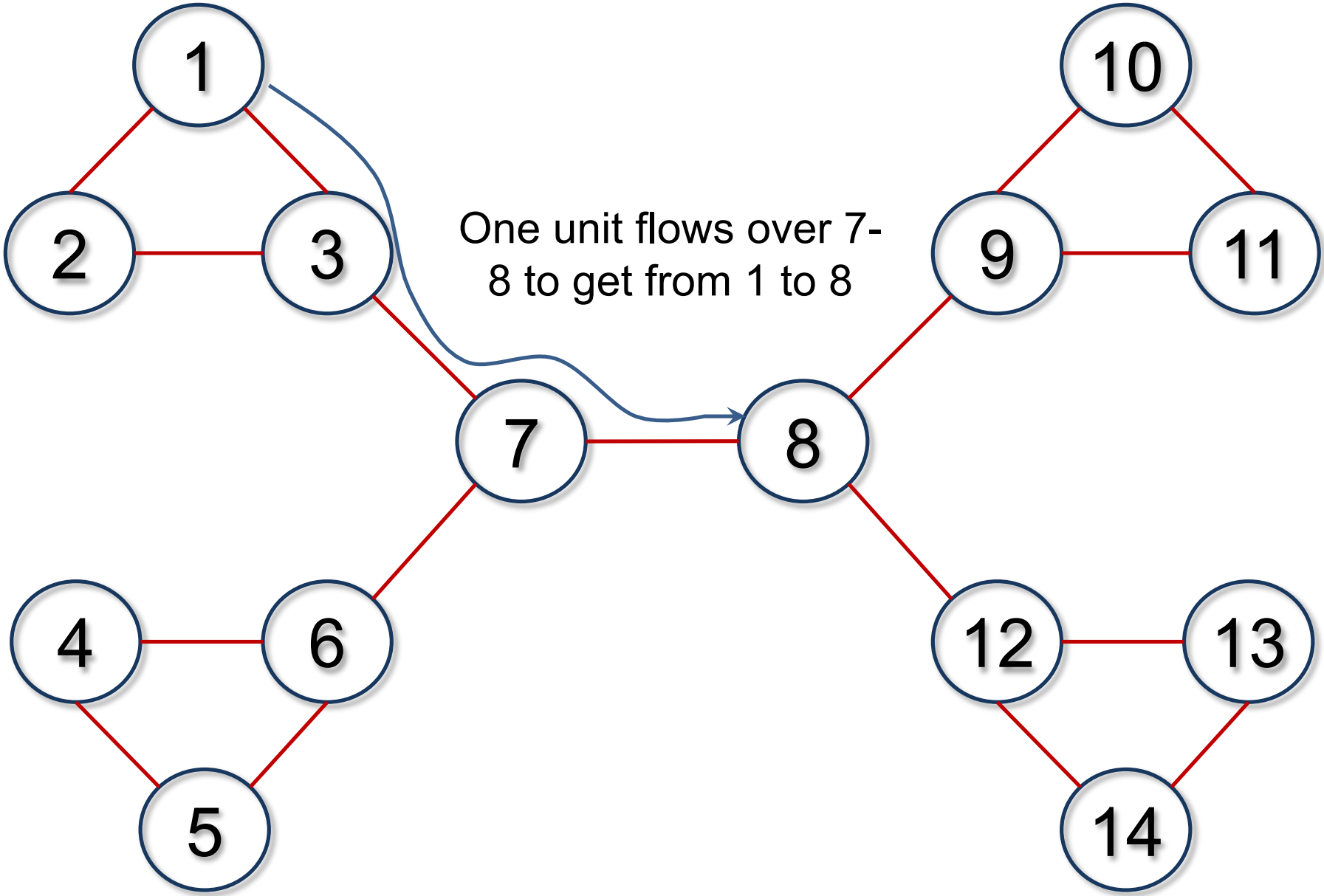


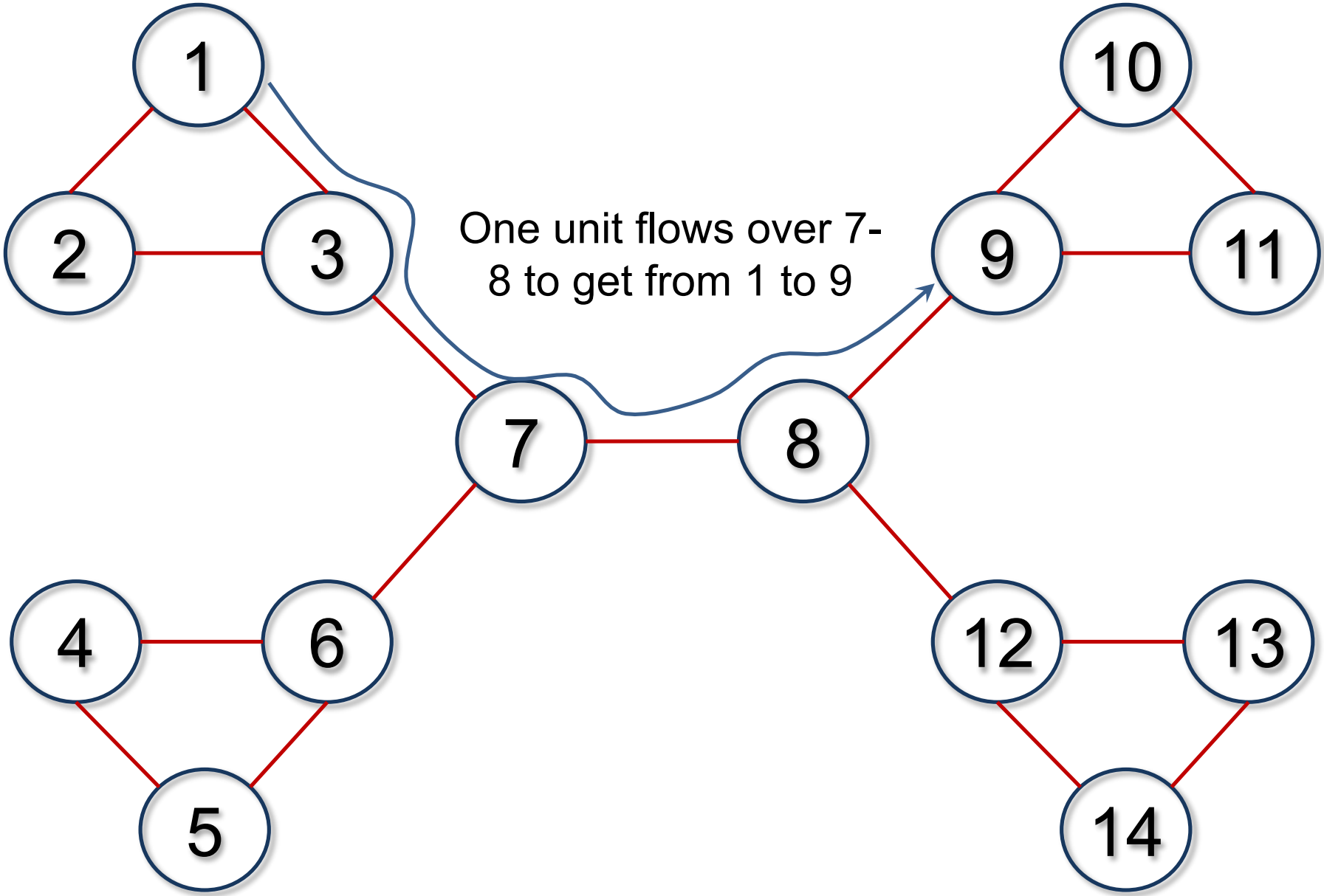



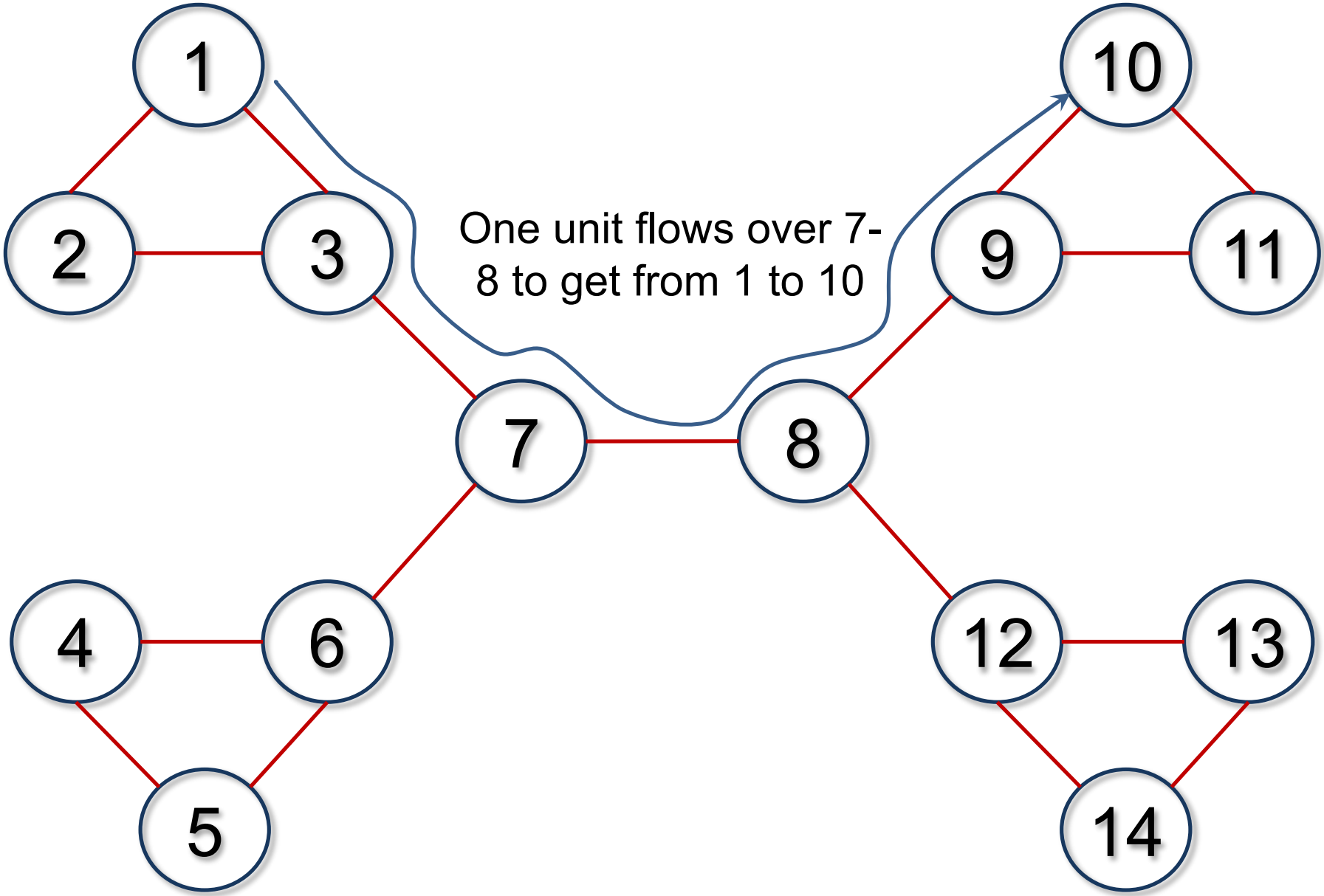

(a) *Step 1* (b) *Step 2*

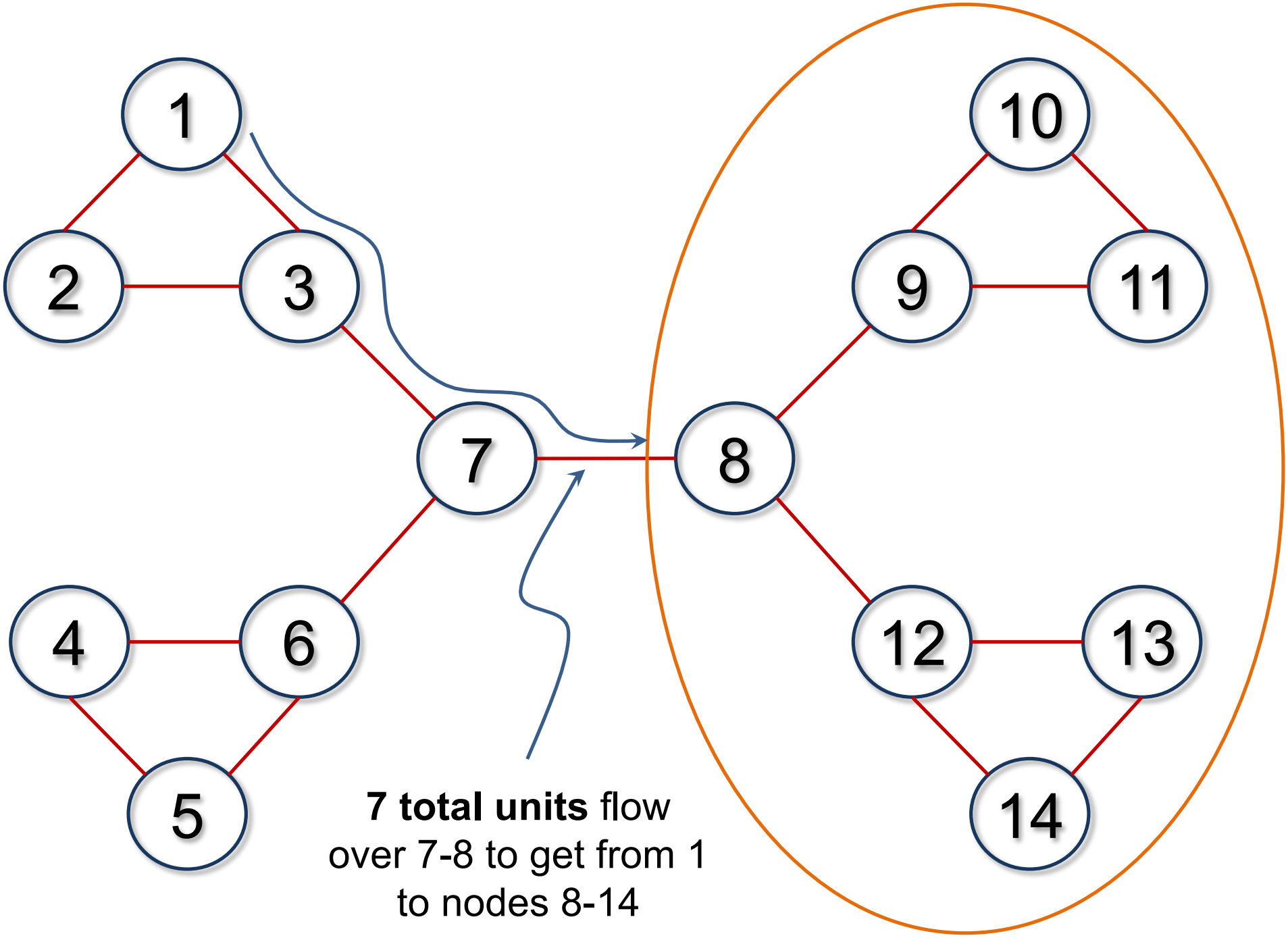
# how do we calculate edge betweenness?

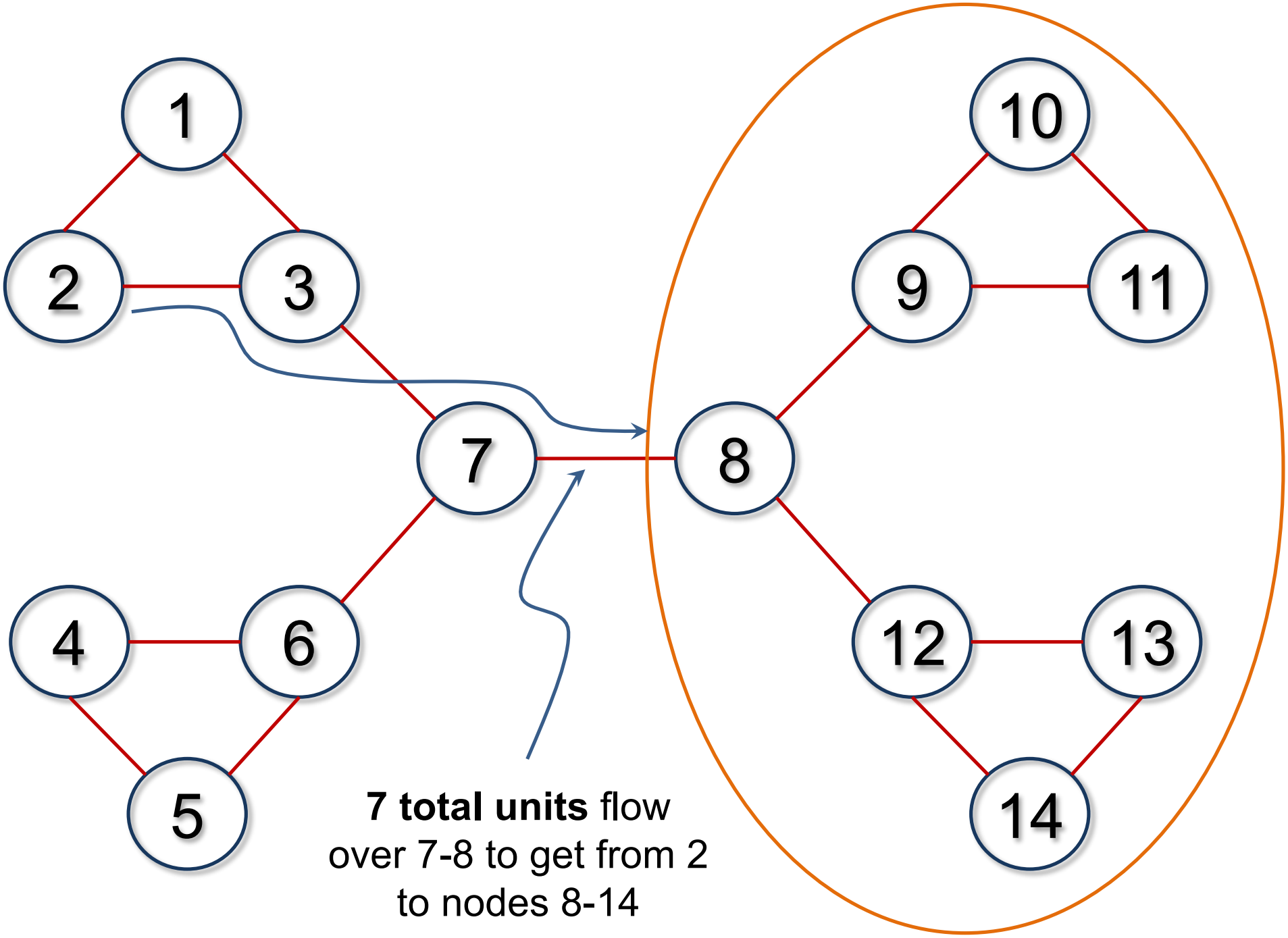


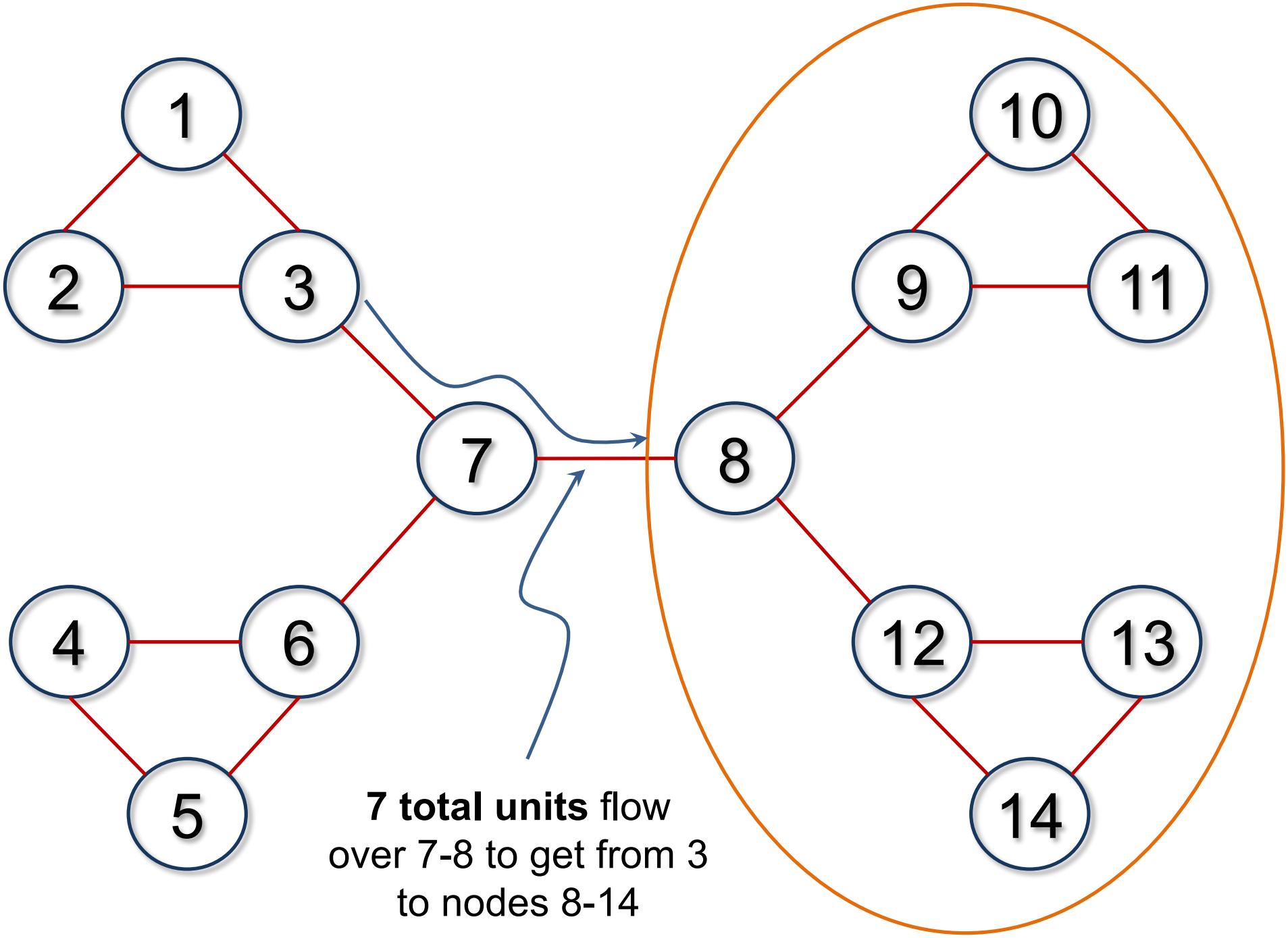

1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8 to get from 1 to 8

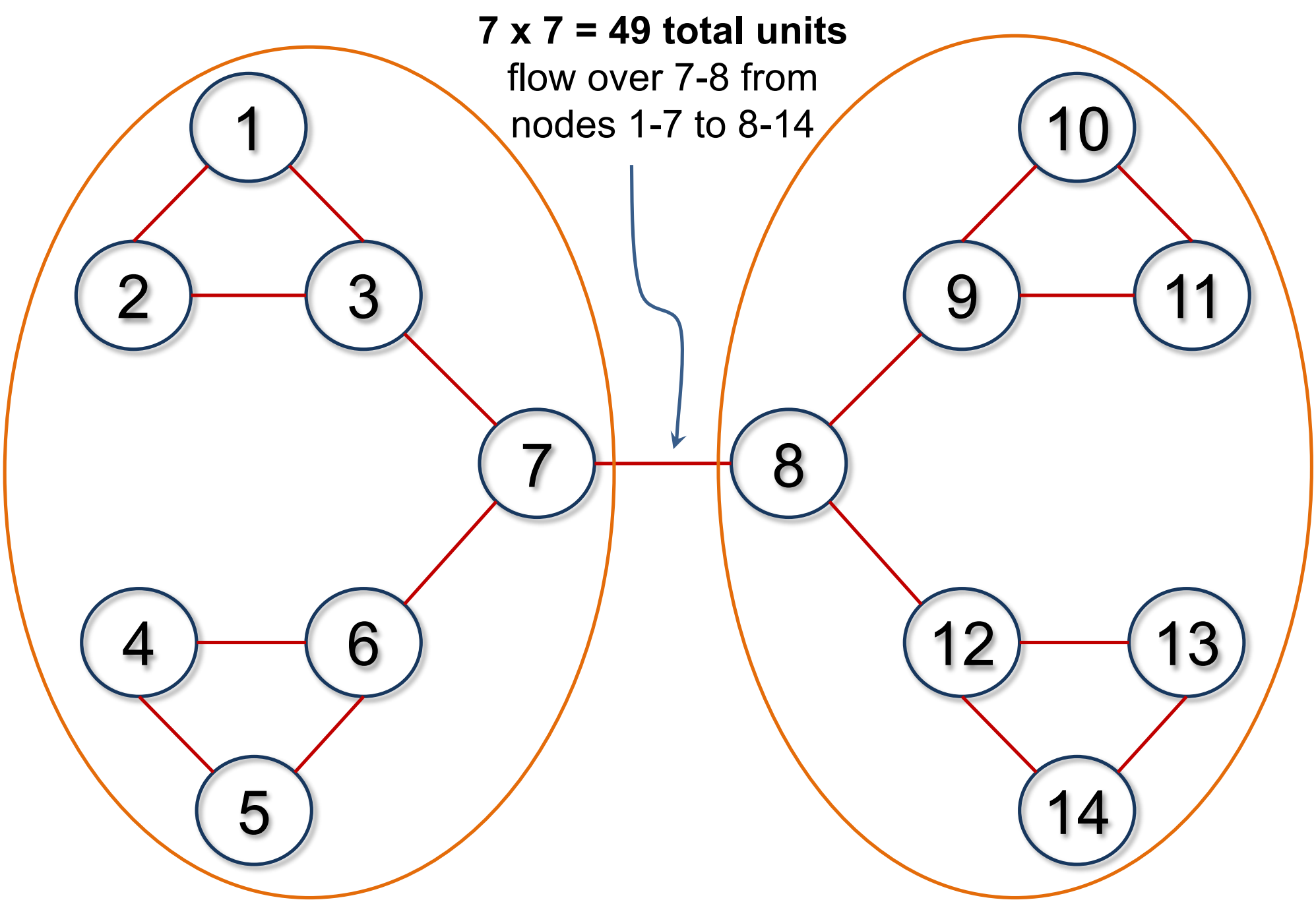

1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8 to get from 1 to 9

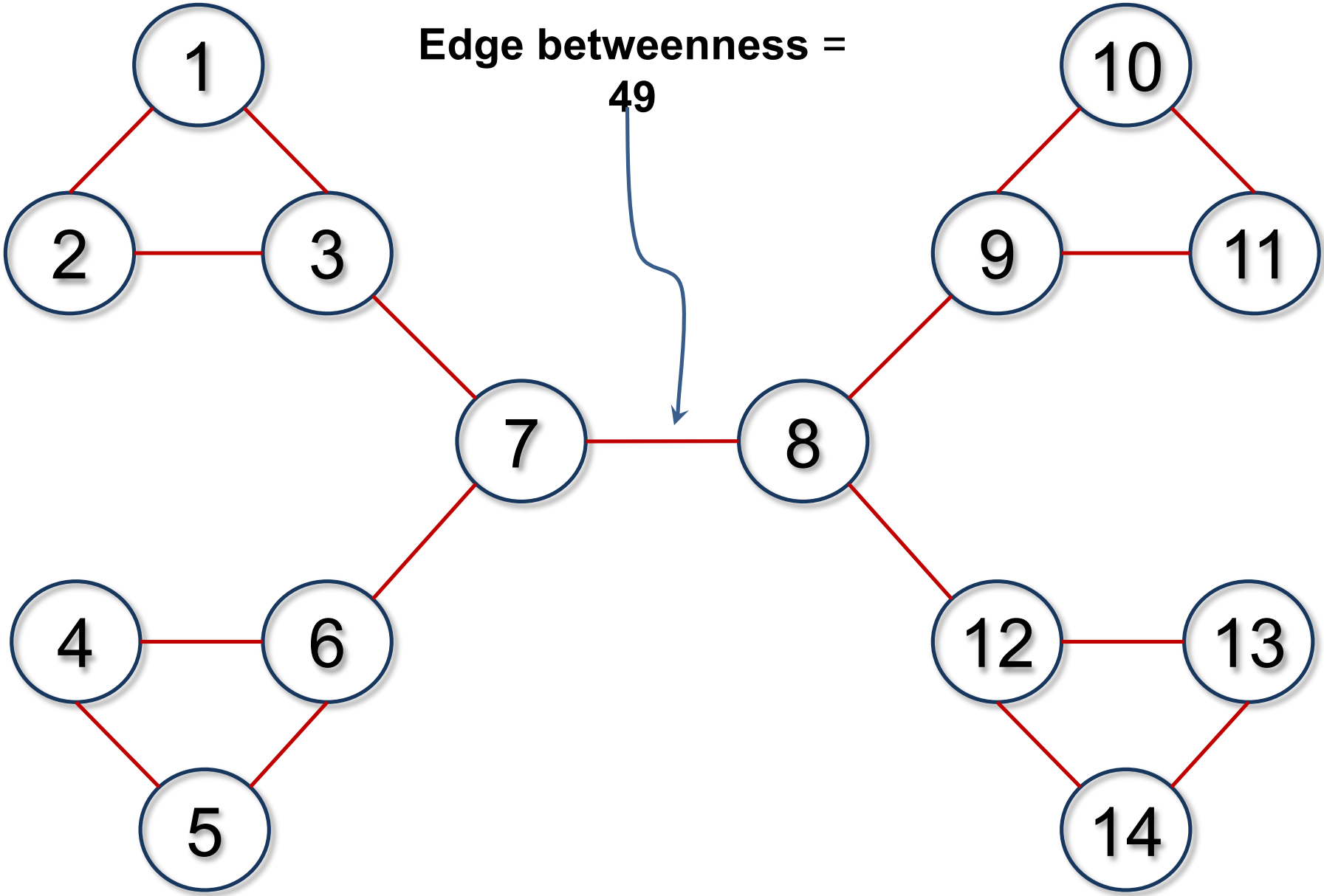

1
2
3
4
5
6
7
8
9
10
11
12
13
14
One unit flows over 7-8 to get from 1 to 10

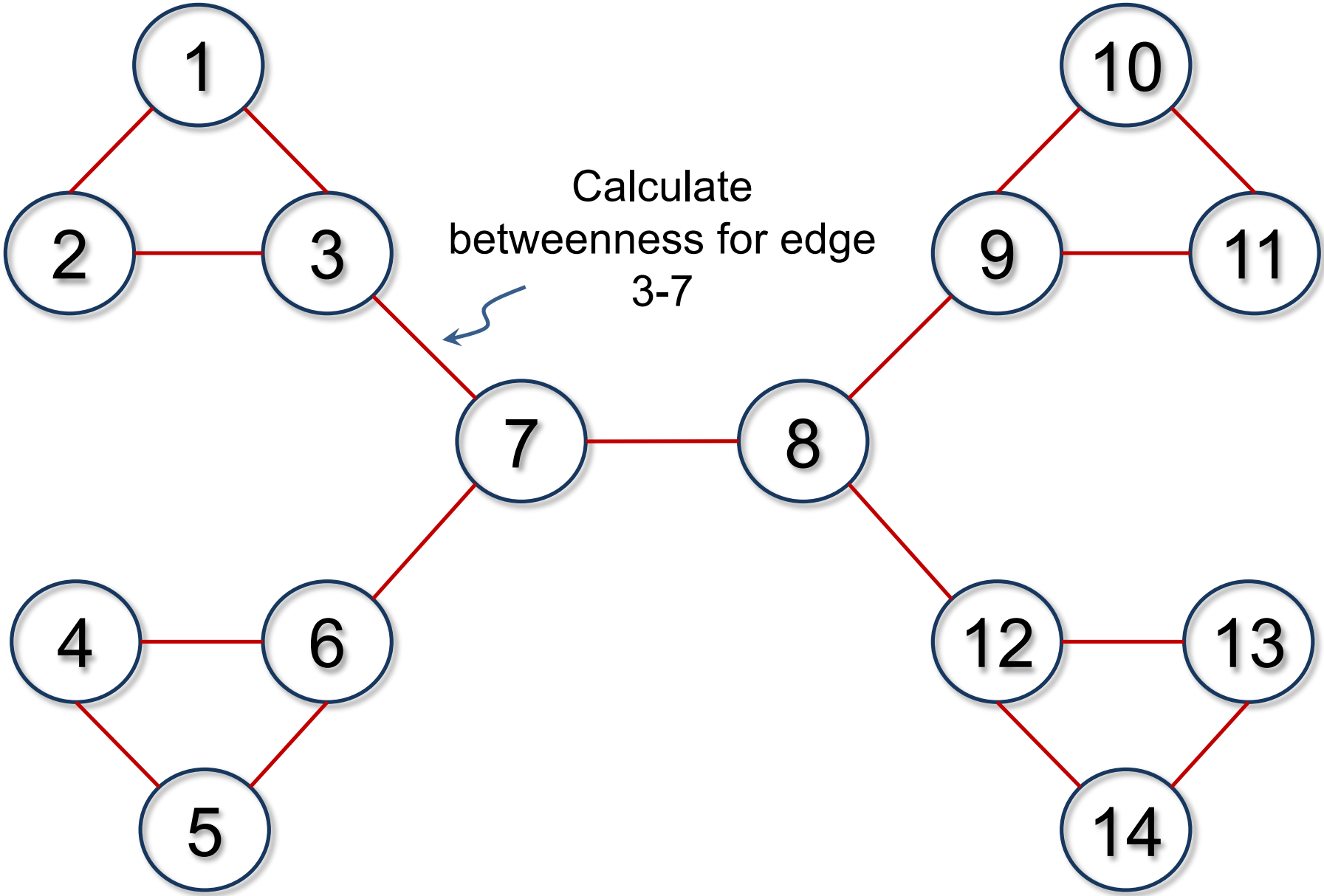

1
2
3
4
5
6
7
8
9
10
11
12
13
14
**7 total units** flow
over 7-8 to get from 1
to nodes 8-14

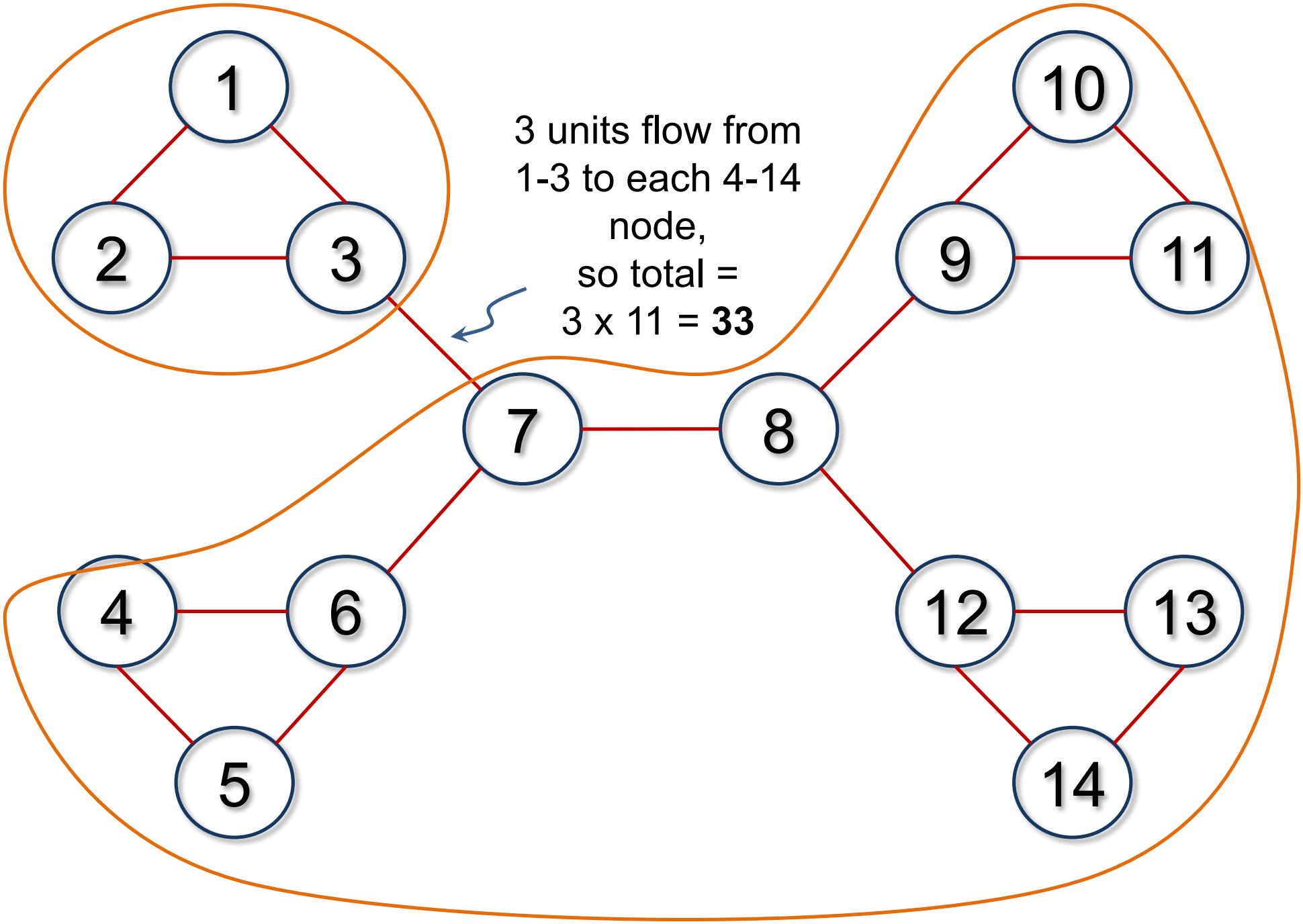

1
2
3
4
5
6
7
8
9
10
11
12
13
14
**7 total units** flow
over 7-8 to get from 2
to nodes 8-14

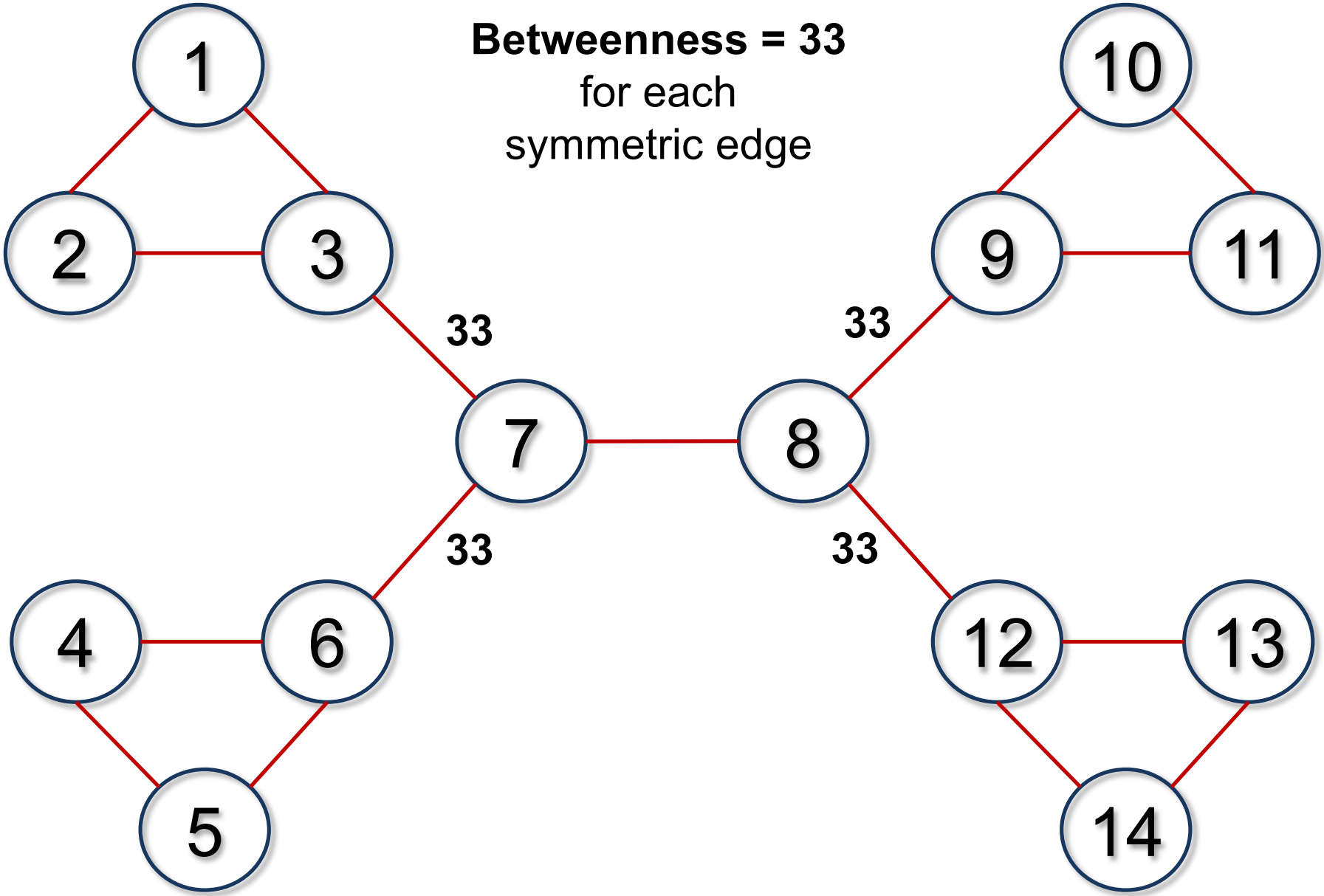

1
2
3
4
5
6
7
8
9
10
11
12
13
14
**7 total units** flow
over 7-8 to get from 3
to nodes 8-14

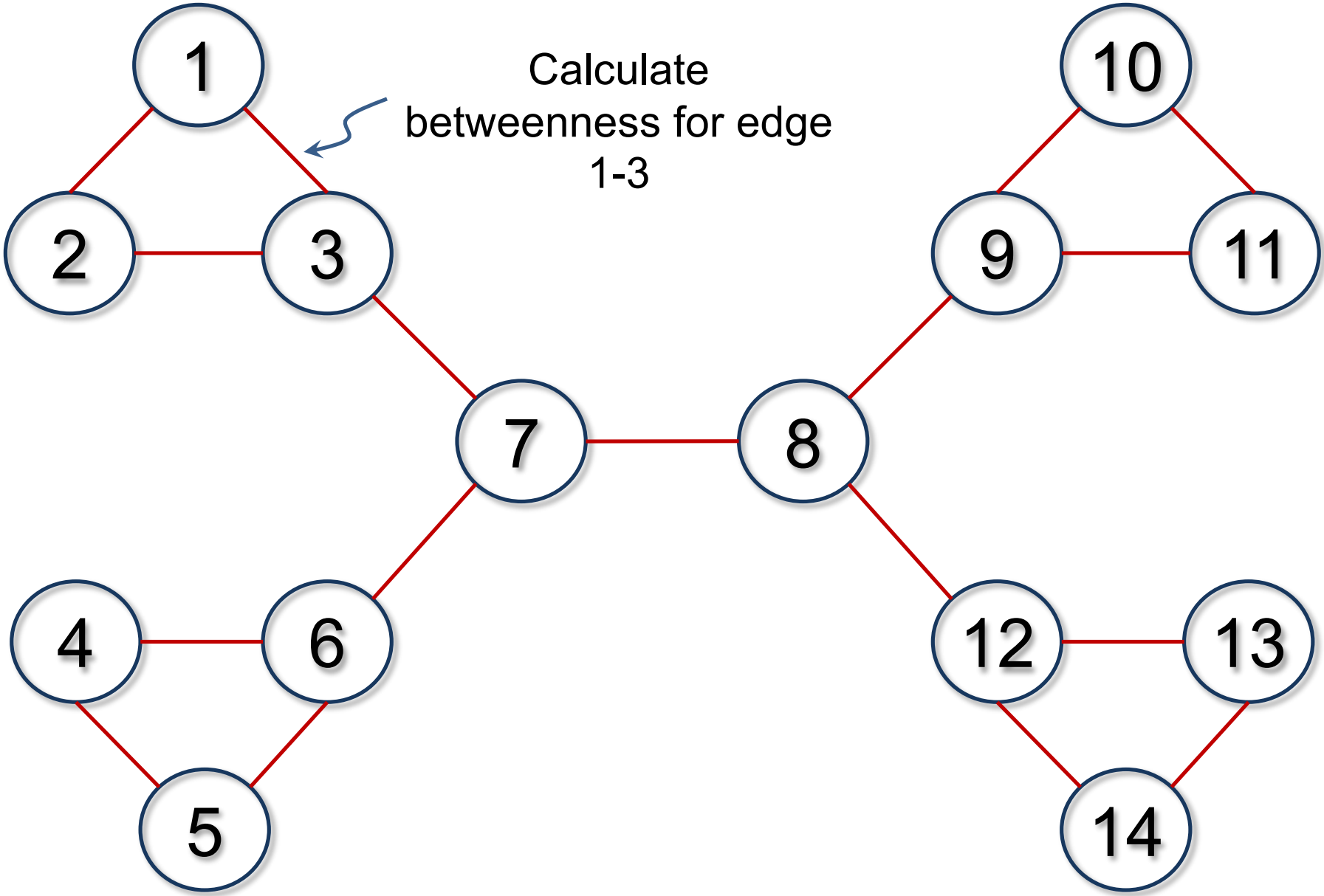

7 x 7 = 49 total units
flow over 7-8 from
nodes 1-7 to 8-14
1
2
3
4
5
6
7
8
9
10
11
12
13
14

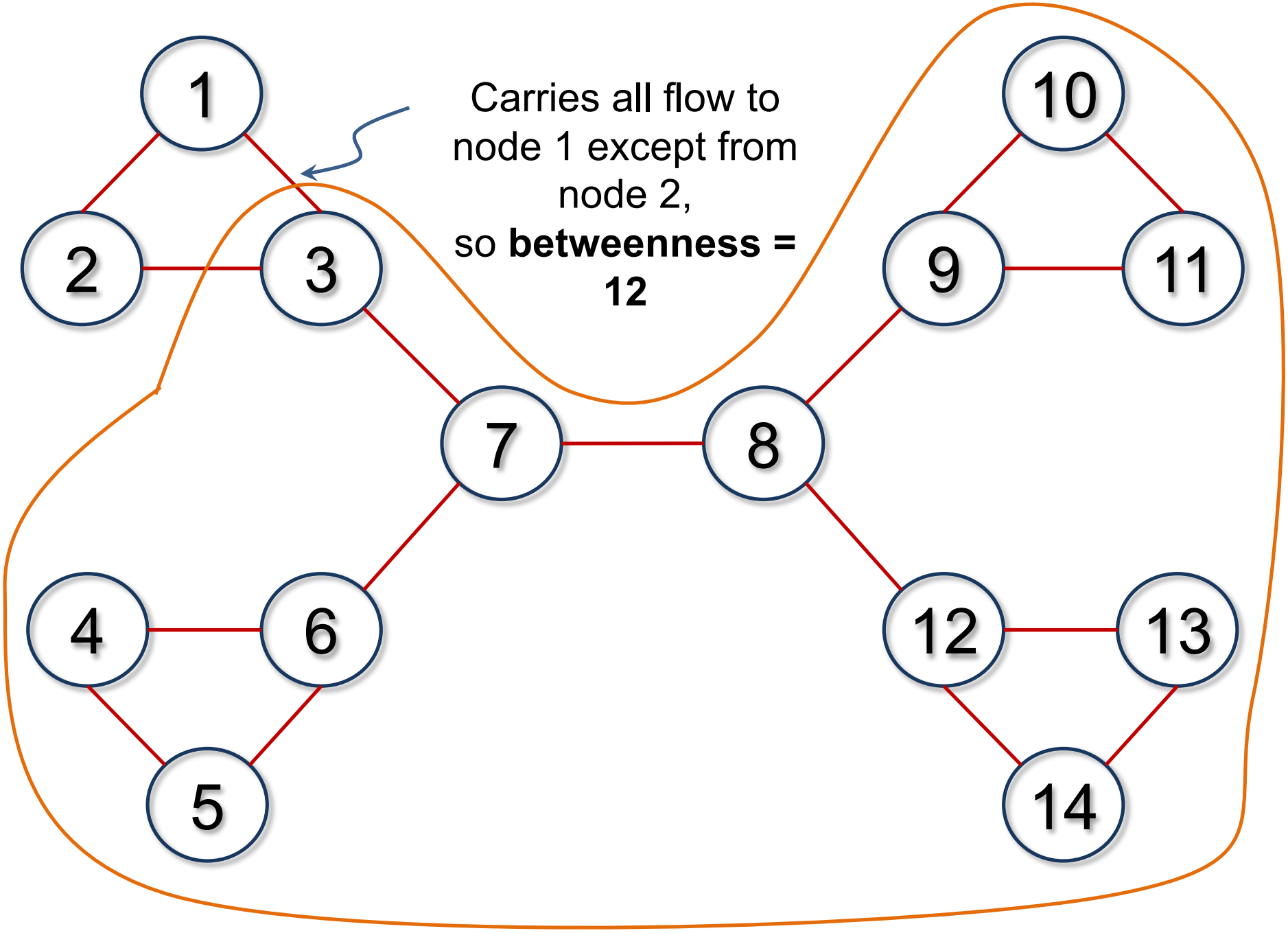

Edge betweenness =
49
1
2
3
4
5
6
7
8
9
10
11
12
13
14

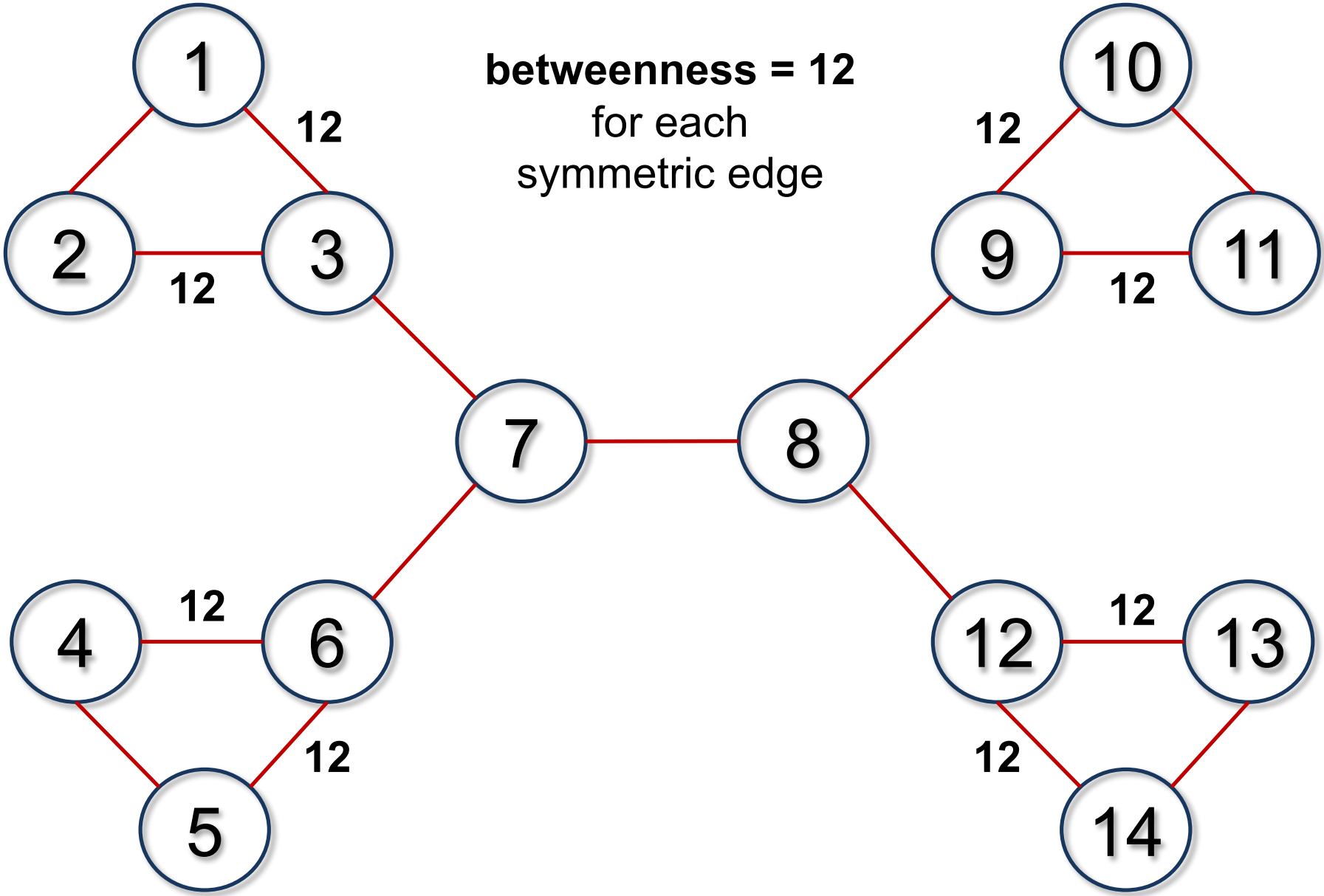

1
2
3
4
5
6
7
8
9
10
11
12
13
14
Calculate
betweenness for edge
3-7

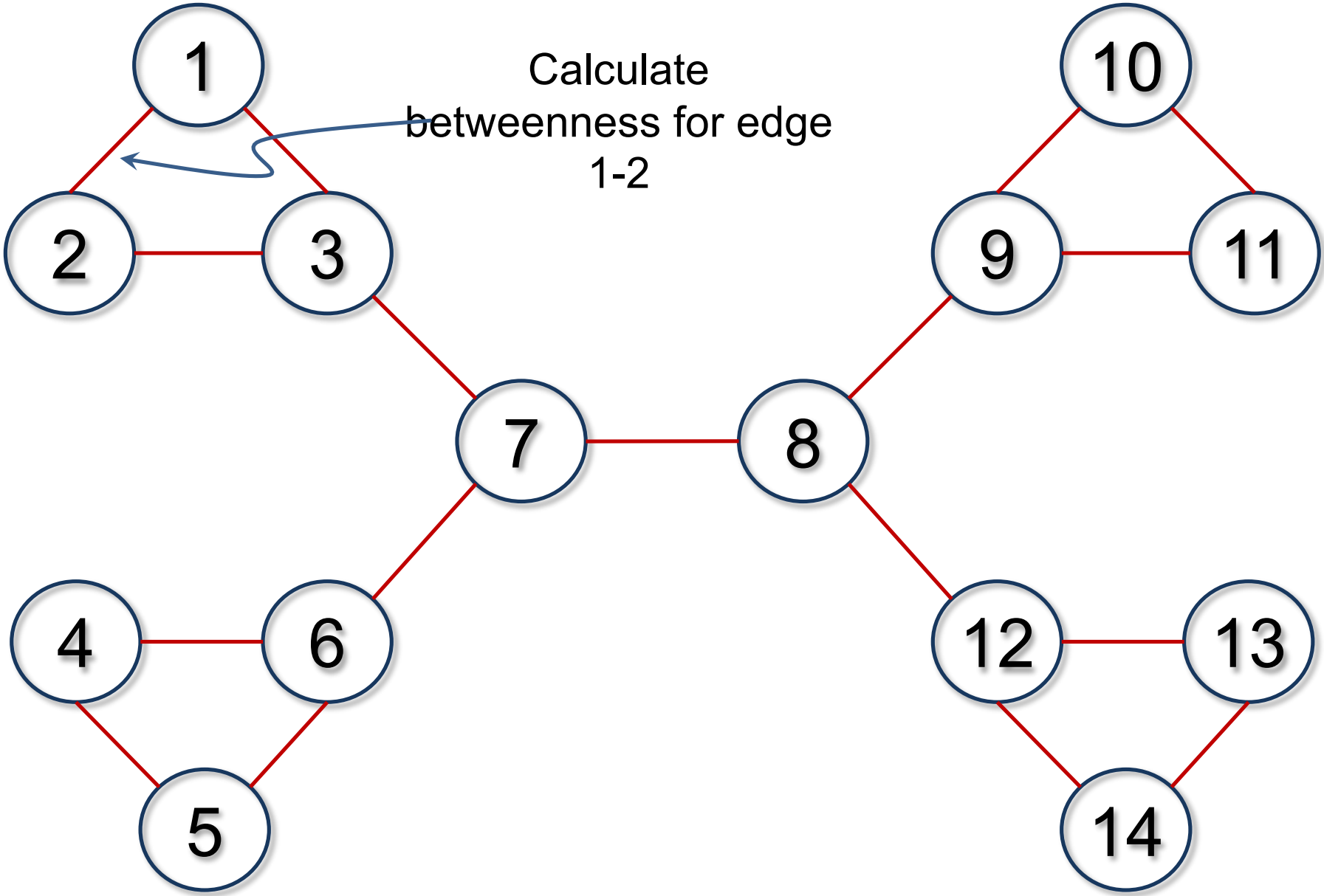

1
2
3
10
9
11
7
8
4
6
12
13
5
14
3 units flow from
1-3 to each 4-14
node,
so total =
3 x 11 = **33**

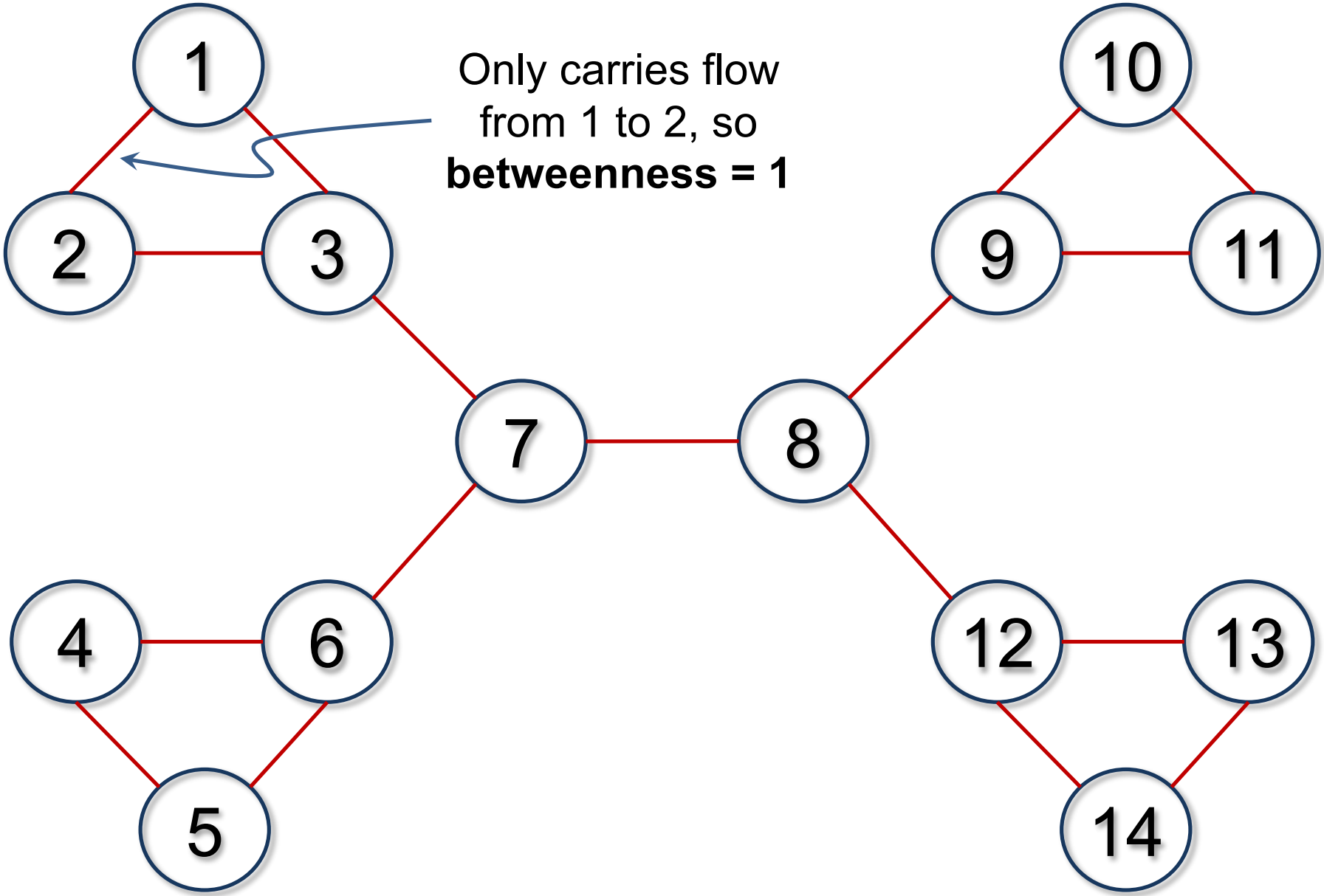

Betweenness = 33
for each
symmetric edge
1
2
3
4
5
6
7
8
9
10
11
12
13
14
33
33
33
33

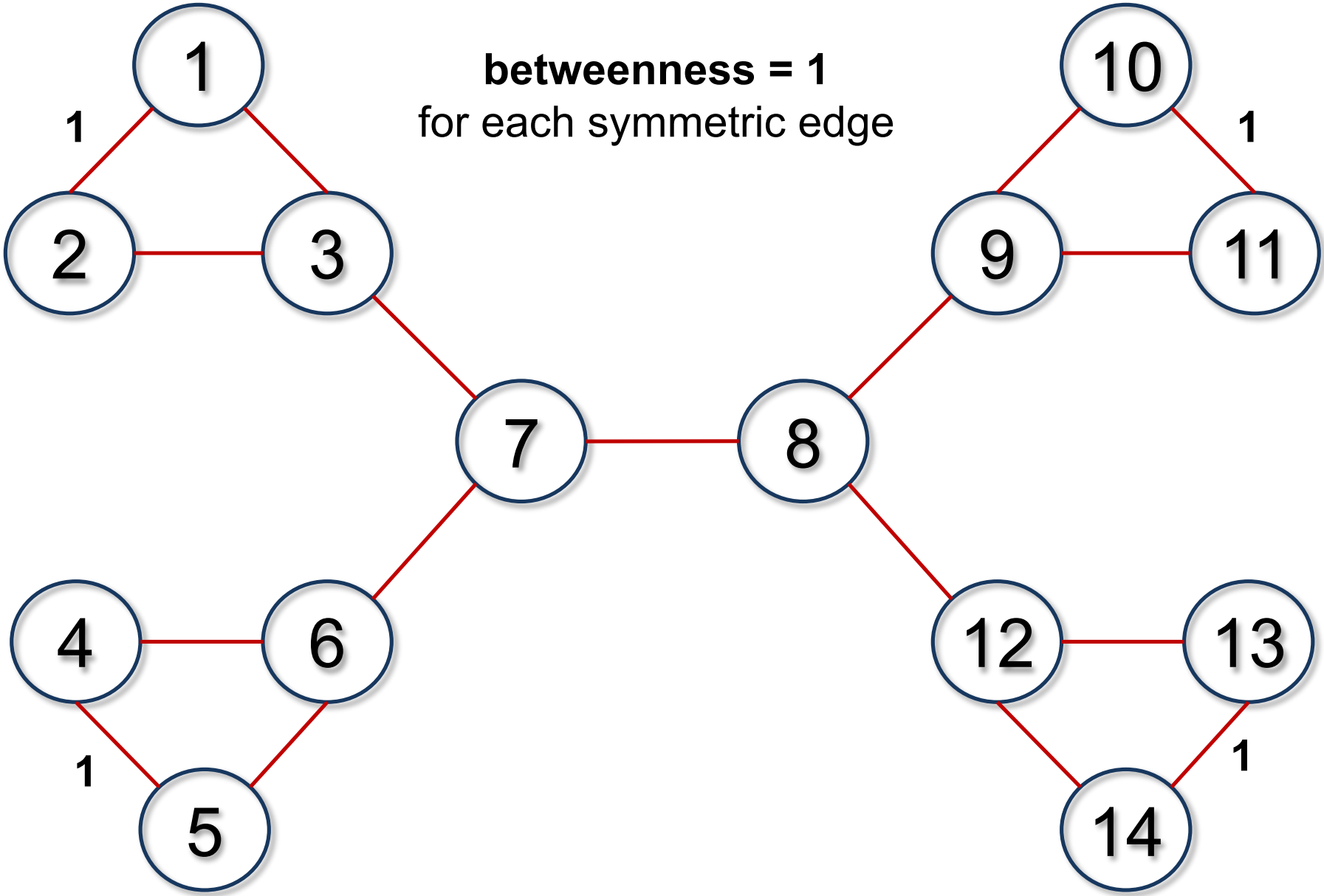

Calculate
betweenness for edge
1-3
1
2
3
4
5
6
7
8
9
10
11
12
13
14

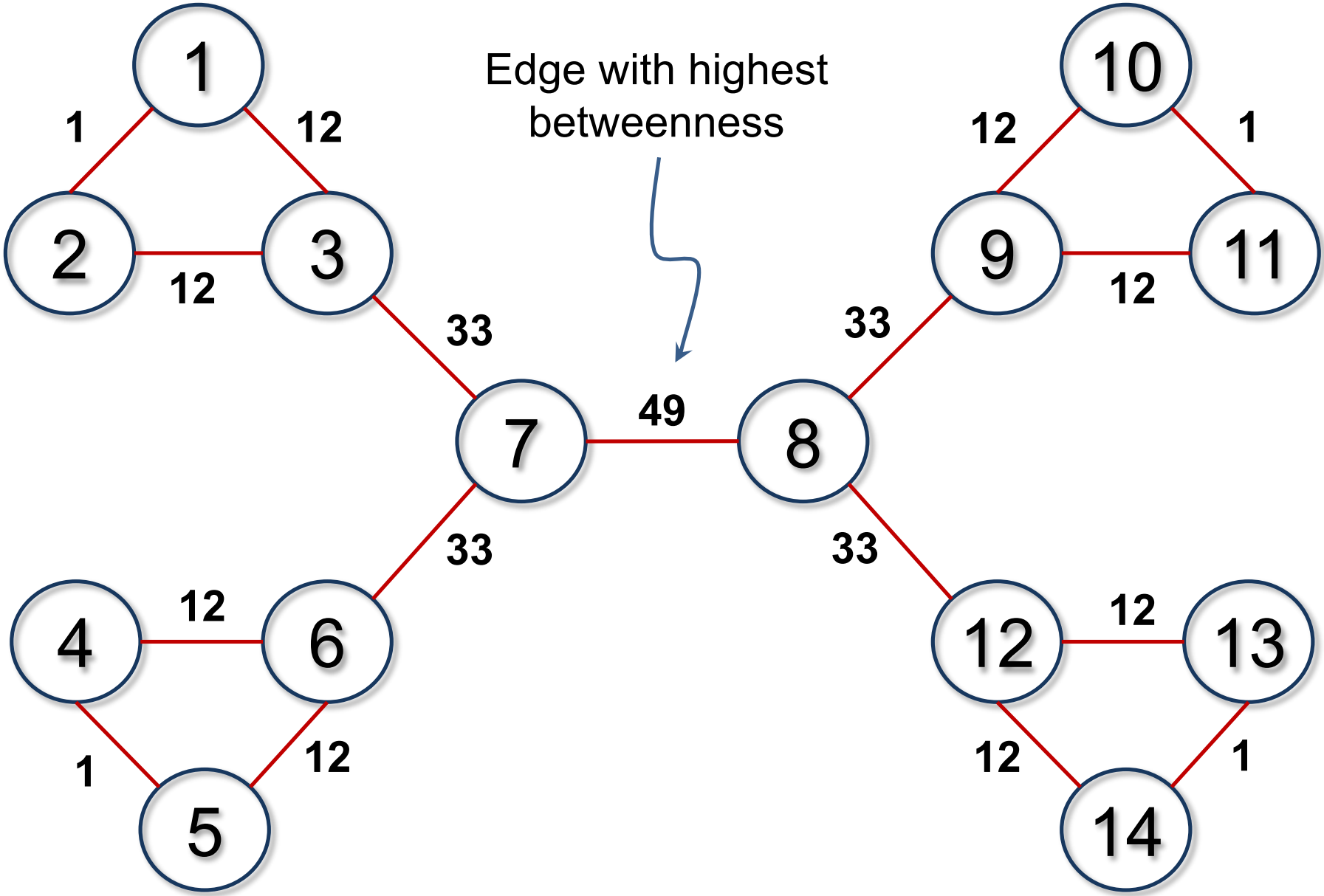

1
2
3
4
5
6
7
8
9
10
11
12
13
14
Carries all flow to node 1 except from node 2, so **betweenness = 12**


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

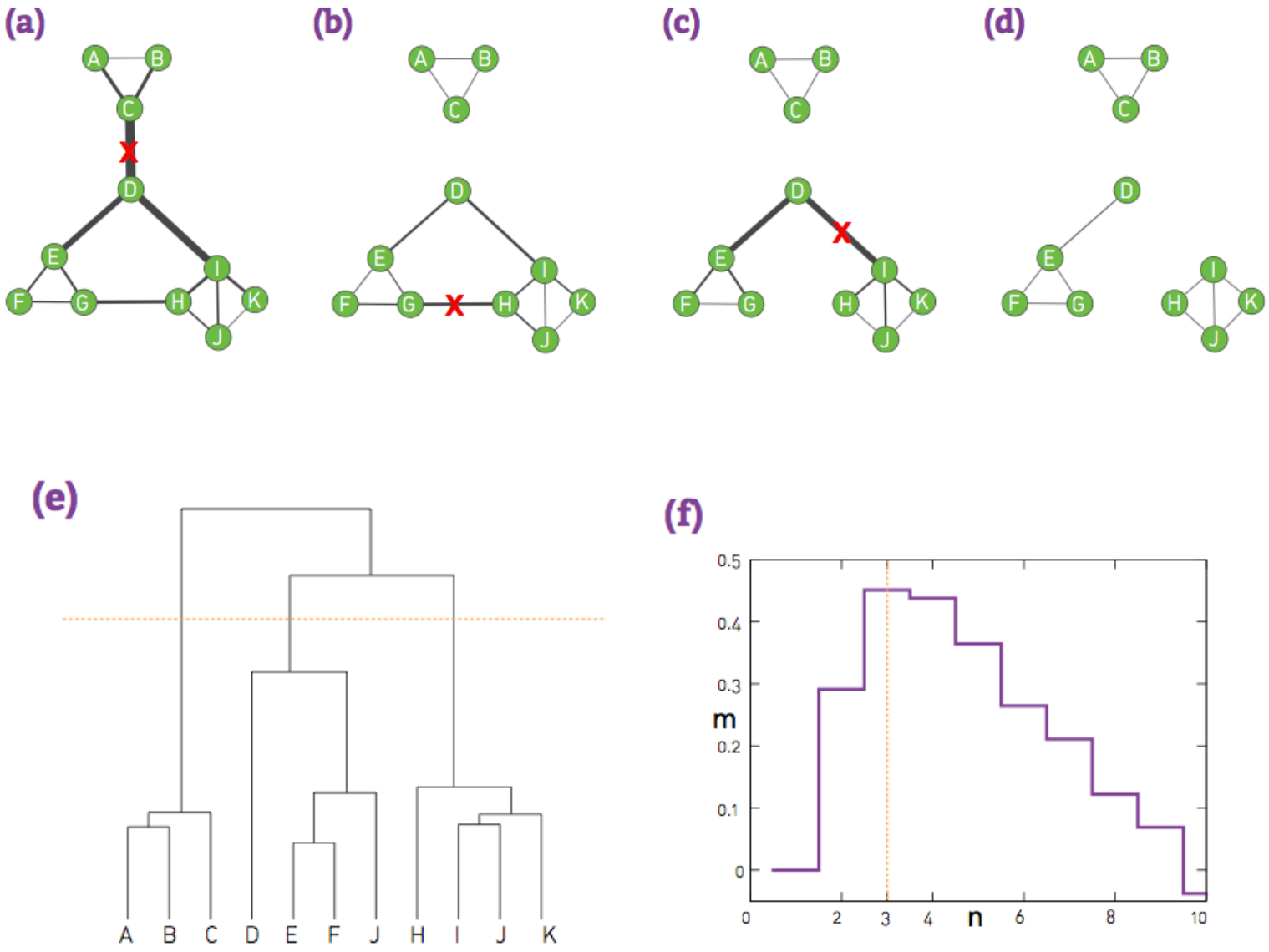

Calculate
betweenness for edge
1-2
1
2
3
4
5
6
7
8
9
10
11
12
13
14


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

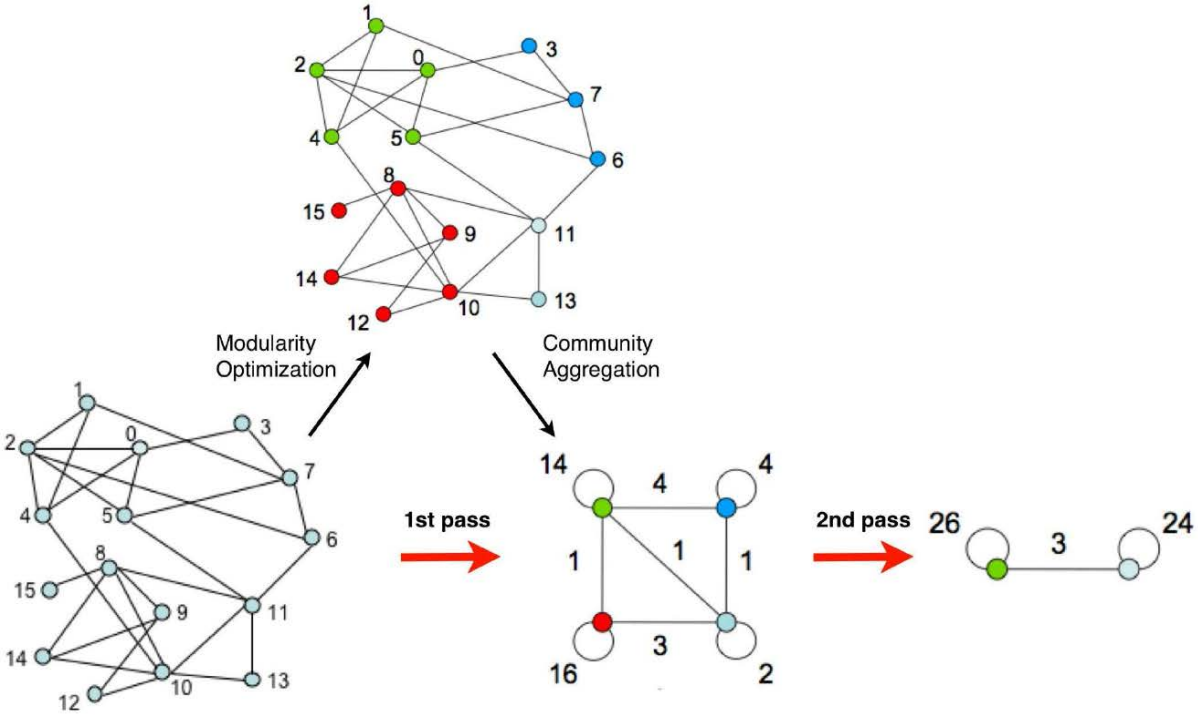

betweenness = 1
for each symmetric edge
1
2
3
4
5
6
7
8
9
10
11
12
13
14
1
1
1
1


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**;



# Louvain method

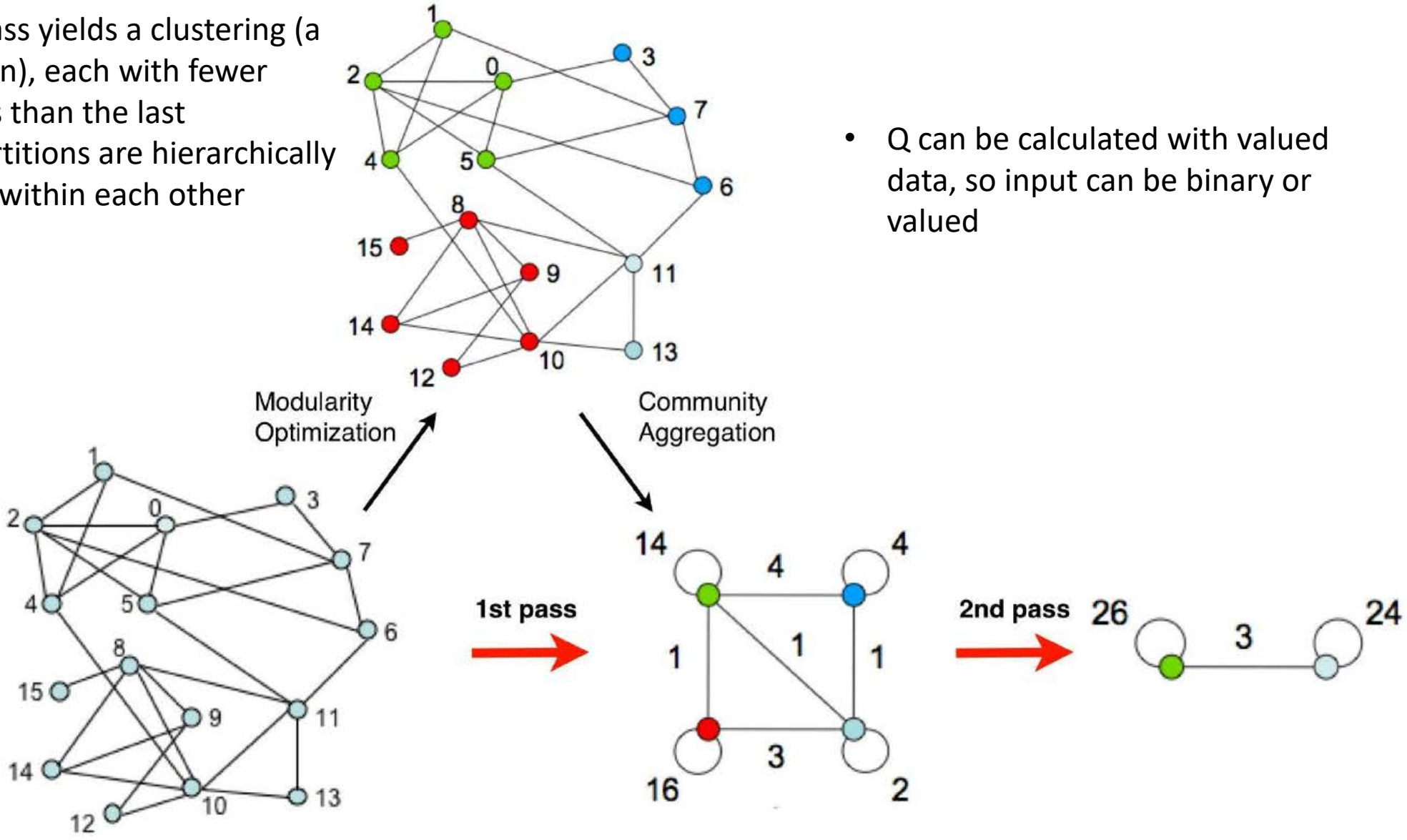



- Start w/ each node in own cluster
- Stage 1
  - For each node i,
    - examine each neighbor j. determine increase in Q or other stat if i were moved to j's cluster
    - Move i to the cluster with the biggest gain
  - Repeat until no further improvements possible
- Stage 2
  - Collapse all nodes within a cluster into a super node, summing all ties to other nodes
- Repeat stages 1 and 2 until no improvement in Q is possible

# Louvain method

->l = Louvain(pv504)
->draw pv504 l

- Each pass yields a clustering (a partition), each with fewer clusters than the last
- The partitions are hierarchically nested within each other




- Q can be calculated with valued data, so input can be binary or valued

# quantifying quality of community structure | Modularity

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

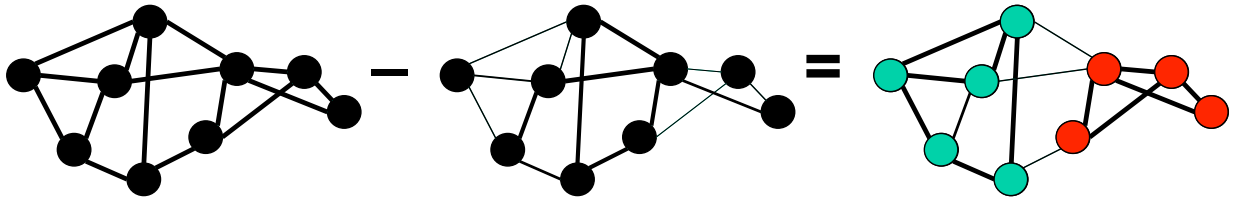
Random graphs are not expected to have community structure, so we will use them as null models.

$Q =$ (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

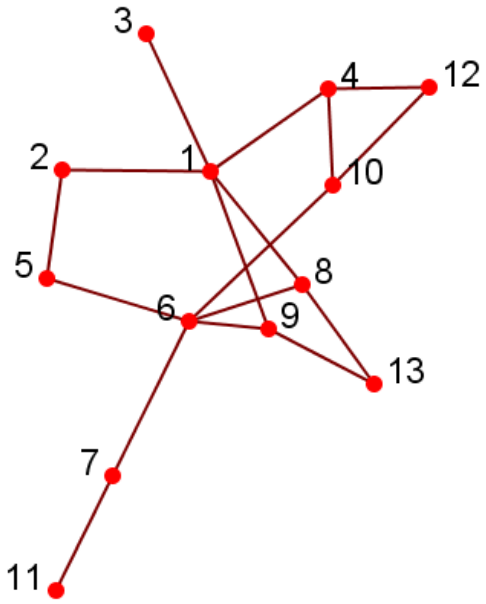
How to computer $\mathbf{P_{ij}}$?

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 $P_{ij} = 2m p_i p_j = \frac{k_i k_j}{2m}$

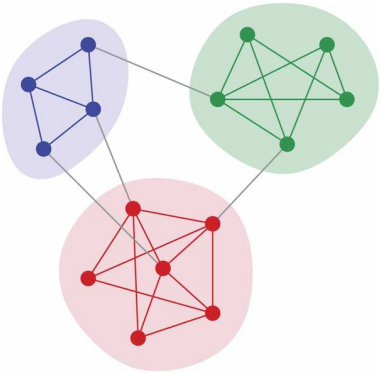



Expected Number of edges between 6 and 9 is
5*3/(2*17) = 15/34

# 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)

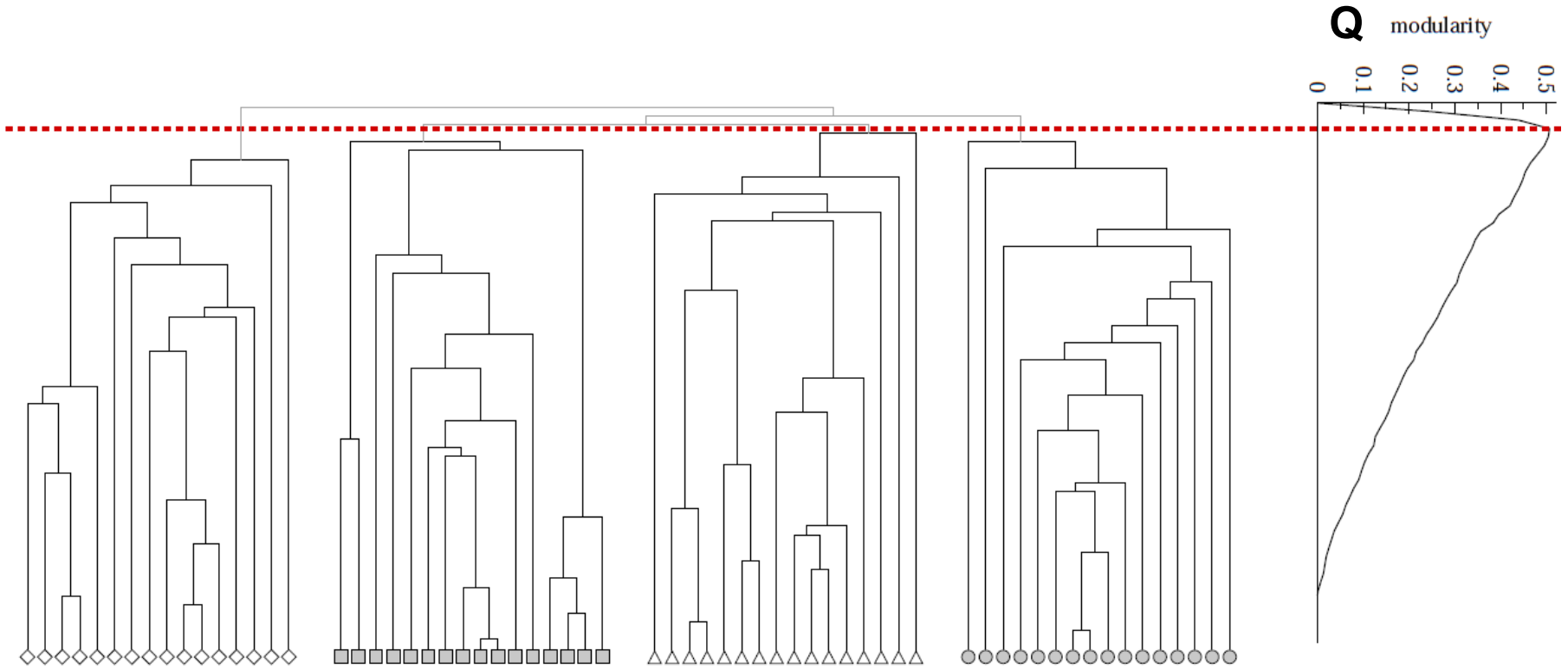
Positive Q means the number of edges within groups exceeds the expected number

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;

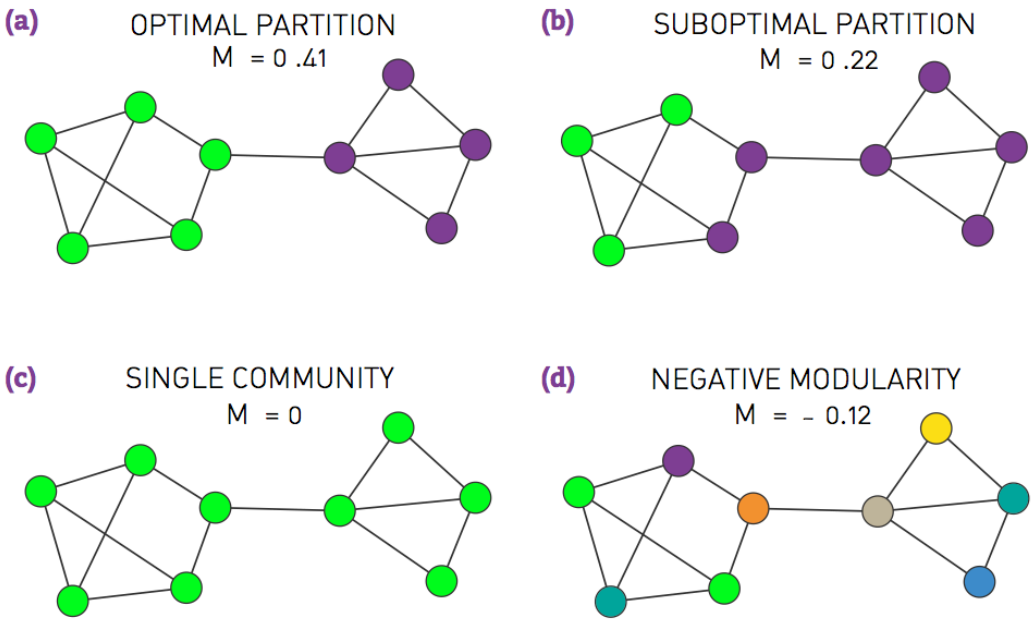
Modularity can be optimized directly (e.g. Louvain algorithm, Spectral algorithm);



# **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 $\Delta M$ 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.

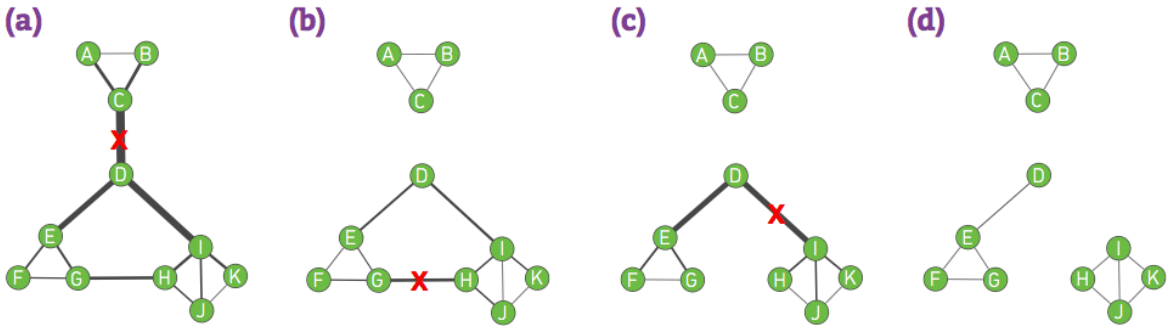
Step 5. Record for each step and select the partition for which the modularity is maximal.

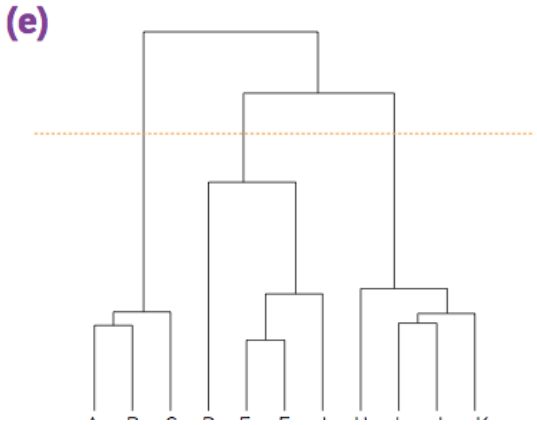
# quantifying quality of community structure | Modularity

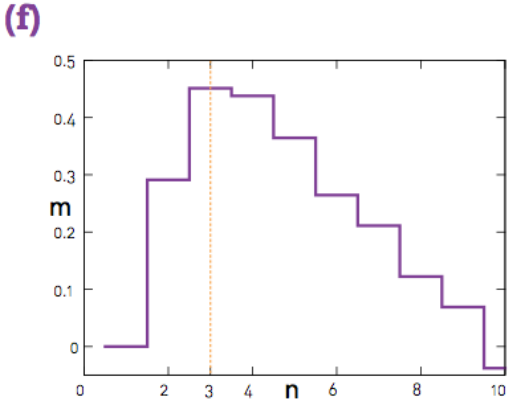
## Optimization

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

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



# Part II - Hypothesis Testing

# Hypothesis Testing with Network Data

# Hypothesis Testing with Network Data

Multiple levels of analysis

| Level | Theory of Networks (network var is Y) | Network Theory (network var is X) |
|---|---|---|
| dyad | For each pair of nodes, predict presence/absence/strength of tie<br>e.g., samesex → friendship<br>Test models of tie formation \| network change \| selection | For each pair of nodes, predict similarity in choices as function of tie between them<br>e.g., years of marriage → similar attitudes<br>Test models of diffusion/contagion/influence |
| node | For each node, predict their centrality<br>e.g., extraversion → number of friends<br><br>Test models of social status attainment | For each node, predict success as a function of social ties<br>e.g., friends in high places → business success<br>Test models of social capital |
| group | For each group, predict the cohesion of network<br>e.g., demographic similarity → density of ties | For each group, predict performance as a function of network structure<br>Structure → function |

# Hypothesis Testing with Network Data

## Two approaches

- **ERGM** -- Exponential random graph models
  - Like a logistic regression predicting presence/absence of tie
  - Handles auto-correlation by explicitly modeling sources of dependency
    - Sender effects like gregariousness
    - Receiver effects like popularity
    - Reciprocity, transitivity
- QAP – Quadratic assignment procedure (permutation test)
  - Like regular regression (or logistic regression) but p-values are constructed by comparing coefs against a distribution calculated from data itself
    - Similar to bootstrapping

# 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

# Intro to permutation tests

- Calculate observed statistic (e.g., corr(X,Y) or difference in means)
- Repeat 10,000 times:
  - Randomly permute values of one variable relative to the others
    - We know these values are independent of the other variable, because they are random permutations
  - Calculate statistic and record whether it was greater than or equal to the observed
- P-value is proportion of times the statistic was greater than or equal to the observed value

Predicting the size of banker's year-end bonus as a function of structural holes in her ego network

| Person | Holes | Bonus | Bonus* |
|---|---|---|---|
| Jim | 3 | 9 | 8 |
| Jen | 9 | 1 | 7 |
| Joe | 2 | 7 | 2 |
| Jill | 1 | 8 | 1 |
| Jon | 15 | 3 | 9 |
| Jeb | 3 | 2 | 3 |

Bonus* is permuted version of Bonus. Holes and Bonus* are causally independent because values of Bonus* were assigned randomly

- A permutation test compares the observed correlation between X and Y against a distribution of correlations obtained by randomly permuting X and Y
- Correlating permuted versions of your variables has two advantages
  - The permuted variables are just like your real variables in every way (e.g., same number of 0s, same average, same std dev, etc)
  - The permuted variables are guaranteed to be independent of your observed data because they were generated randomly

# 1. Dyadic Hypotheses

->unpack Padgett
->qap padgm padgb

## Permutation tests for dyadic variables (QAP)

- Re-order rows and corresponding columns of the matrices in order to produce new dyadic variables that have same constraints as real variables but are necessarily independent

| | jim | jill | jen | joe |
|---|---|---|---|---|
| jim | 0 | 50 | 61 | 57 |
| jill | 50 | 0 | 85 | 41 |
| jen | 61 | 85 | 0 | 54 |
| joe | 57 | 41 | 54 | 0 |

| | jen | jill | jim | joe |
|---|---|---|---|---|
| jen | 0 | 85 | 61 | 54 |
| jill | 85 | 0 | 50 | 41 |
| jim | 61 | 50 | 0 | 57 |
| joe | 54 | 41 | 57 | 0 |

No triadic dependencies are broken when permuting in this way

- Call this approach QAP correlation (and QAP regression, etc)
  - Correlate matrices (this is the observed test statistic)
  - Permute rows/cols of one matrix. Re-correlate. Repeat 10,000 times
  - P-value is the proportion of correlations that are as large as the observed

# 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 |

+

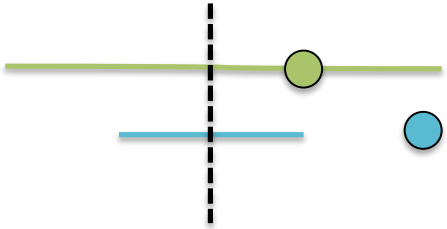
| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 0 | 1 | 0 | 2 | 1 | 0 | 0 |
| B | 1 | 0 | 3 | 5 | 1 | 4 | 2 |
| C | 0 | 3 | 0 | 4 | 5 | 8 | 10 |
| D | 2 | 5 | 4 | 0 | 0 | 3 | 2 |
| E | 1 | 1 | 3 | 0 | 0 | 2 | 2 |
| F | 0 | 4 | 2 | 3 | 3 | 0 | 1 |
| G | 0 | 2 | 1 | 2 | 2 | 1 |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

Friendship tie

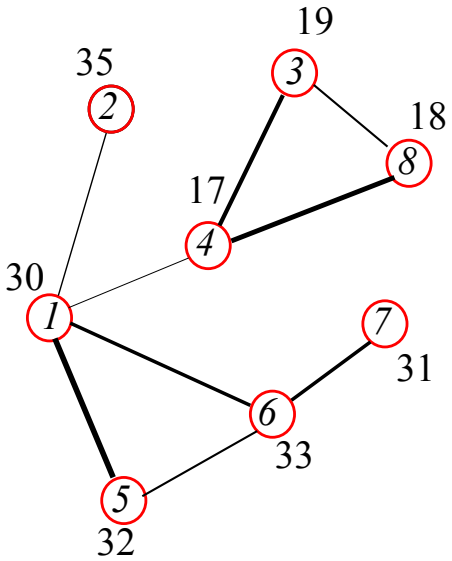
Age difference

education

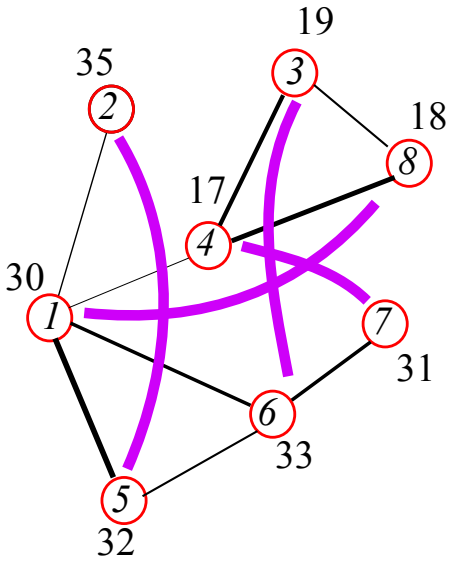
- Permutes dependent variables lots of time. Measure the sampling distribution of the coefficients.
- P-value is a proportion of times that the observation is Falling outside the sampling distribution.



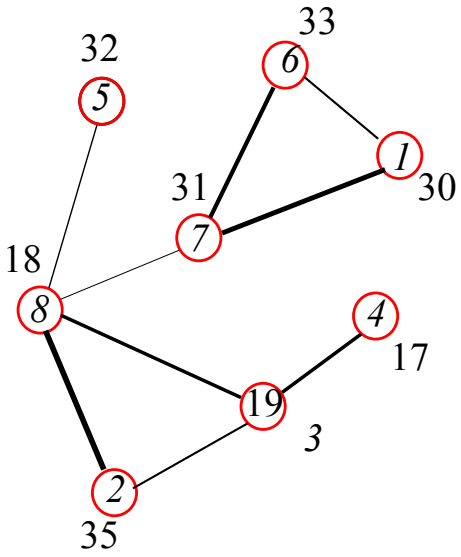
# QAP process – graph representation




before




reshuffling




after

- Unpack krack-high-tec
- Press Ctrl-R for regression

# QAP regression (MR-QAP)

- Predicting advice-seeking as a function of being friends with that person and controlling for reporting to that person
  - Advice(i,j) = b0 + b1*friendship(i,j) + b2*reports_to(i,j)

# MRQAP

- The MRQAP approach was developed by Hubert (1987) and Krackhardt (1987, 1988).
- The basic idea is to apply regular regression coefficients and OLS linear regression analysis to dyadic data collected in square matrices;
- compute *p*-values by a *permutational approach*:
  - the null distribution is obtained by permuting *X* values and *Y* values with respect to each other, permuting rows and columns ('actors') simultaneously so that the network structure is respected.
- This does not model network structure, but controls for it.
- The MRQAP approach is especially useful if one is not interested in network structure per se, but wishes to study linear relations between dyadic independent and dependent variables in a network setting.

# MRQAP – cont.

- It was shown by Dekker, Krackhardt and Snijders (2007) how to do this correctly when controlling for other variables (permute residuals; use pivotal statistics).
  - In ucinet this is called the “double dekker” method
- For each X variable X(k),
  - Regress X(k) on all other X variables. Construct the residual matrix R(k)
  - Regress Y on R(k) together with all the other X variables
    - the beta b(k) on R(k) is the observed beta. It is same value as you would obtain if you simply regress Y on all of the X variables
    - Repeat 10,000 times, permuting rows/cols of R(k)
    - Count the proportion of random permutations that yield a value b(k) as large as the observed b(k)
  - The Xs participate in two regressions, hence the “double” part of the name

# MR-QAP via Double Semi-Partialling

- Dekker, Krackhardt and Snijders (2007) how to do this correctly when controlling for other variables (permute residuals; use pivotal statistics).
- Suppose we want to see effect of X on Y controlling for Z
  - Y = b0 +b1X + b2Z
- Model X as a function of Z and construct residuals
  - X = m0 + m1Z
  - Xres = X – (m0 + m1Z)
- Model  Y as a function of both Xres and Z
  - Y = b0 + b1Xres + b2Z
- Permute rows and columns of Xres 10,000 times and rerun the regression. Calculate t statistic for b1 and count how often the observed t is greater than or equal to the t value in the permuted data
  - For 2-tailed test do abs(t) >= abs(t for π(Xres))
- Z is partialled out twice, hence the name double semi partialling or double dekker
- T-statistic is example of a pivotal statistic. This is as important as the double partialling

# Some dyadic hyps are actually cross-level

- Selection example (homophily/heterophily)
  - Node attribute: gender
  - Dyadic tie: whether i and j meet at conference
  - Sample hypotheses
    - Homophily. People seek out similar others to talk to, make friends with etc
    - Appeal. Women are easier to talk to, so both men and women seek out women
- Influence example (diffusion, contagion, learning)
  - Node attribute: eating octopus
  - Dyadic tie: amount of interaction
  - Sample hypotheses
    - Pressure/modeling behavior. Friends eat octopus, so it becomes thinkable, normal
    - Revulsion. Friends eat octopus in front of you. You decide you will never do that …

# 2. 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

# 3. 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

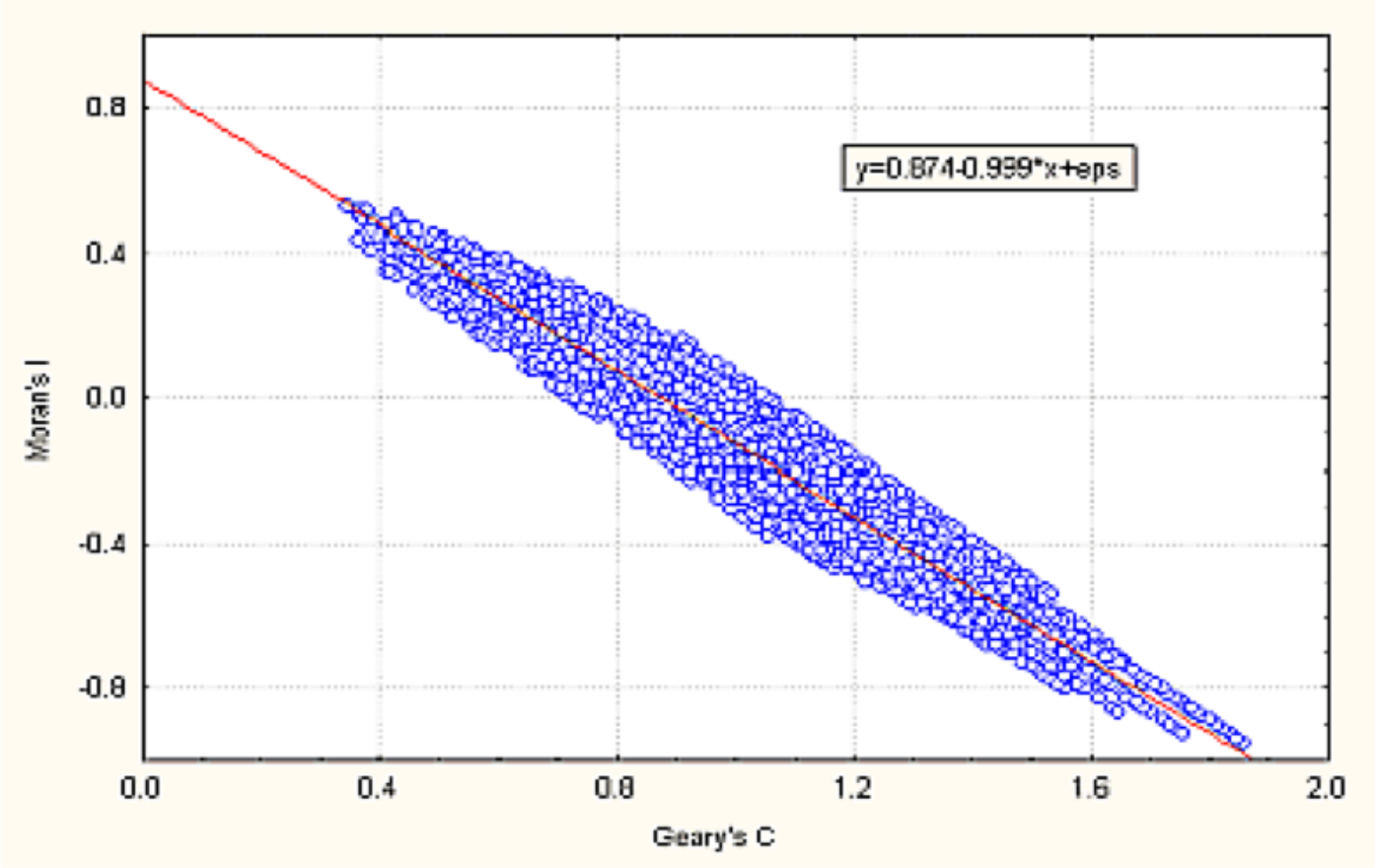
# Continuous Autocorrelation

- 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

- [classically dealt with as spatial autocorrelation (drawn from geography]
- 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

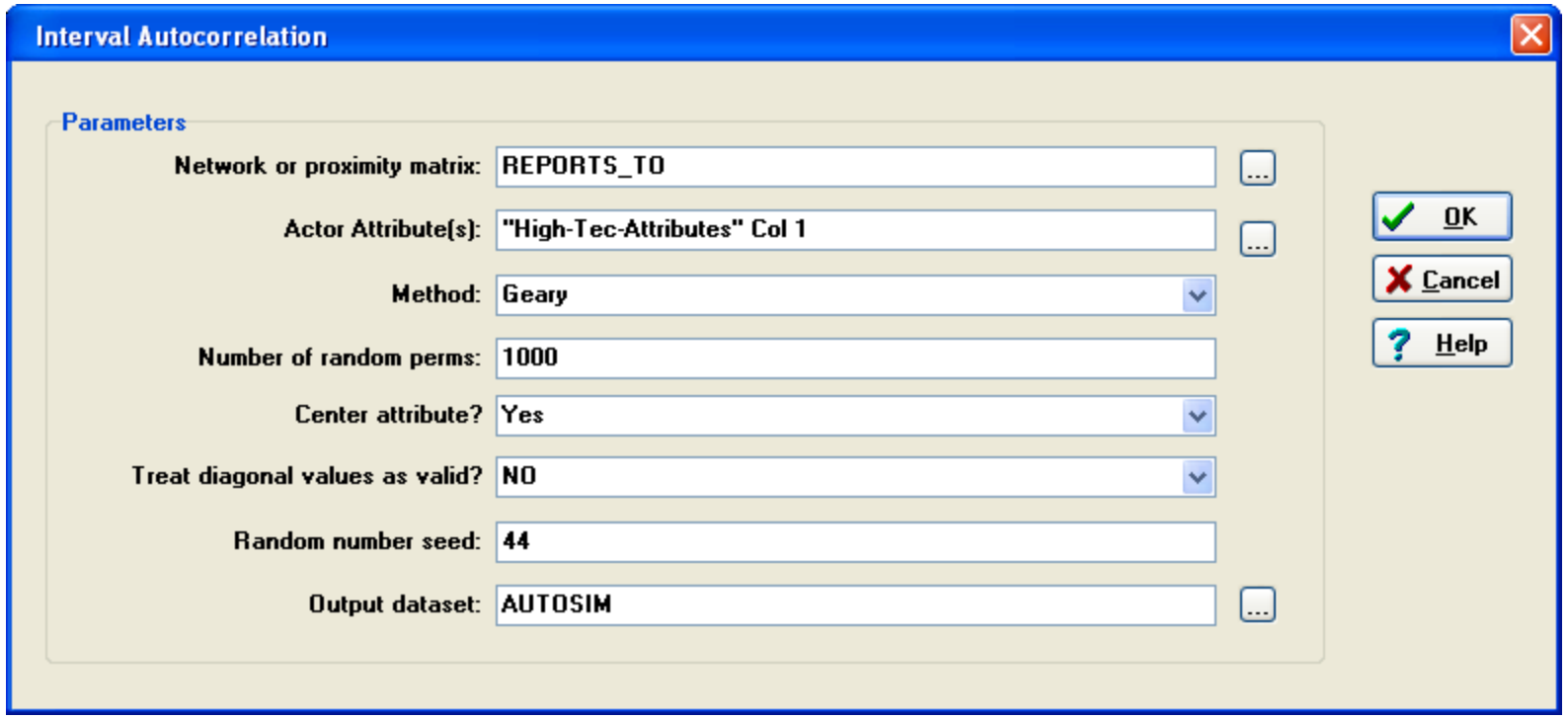
- Let $w_{ij} > 0$ indicate adjacency of nodes i and j, and $X_i$ indicate the score of node i on attribute X (e.g., age)

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

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

# Krack High Tec

Do people report to those of a different age ie negative autocorrelation



```
Method:                                    Geary
# of Permutations:                        1000
Center attribute?                          YES
Random seed:                               44

NOTE: Smaller values indicate positive autocorrelation.
      A value of 1.0 indicates perfect independence.

          Autocorrelation:       0.814
            Significance:        0.385

      Permutation average:       0.986
           Standard error:       0.357
      Proportion as large:       0.615
      Proportion as small:       0.385

----------------------------------------
```

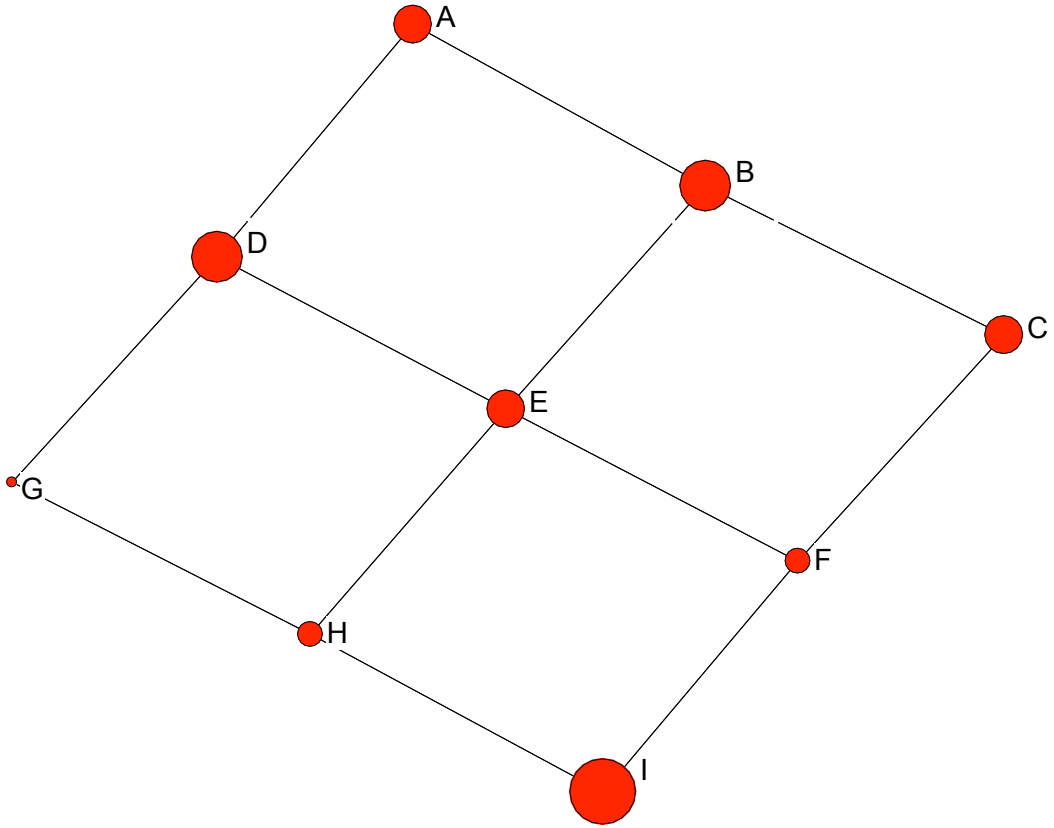
# Moran's I

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

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

# No Autocorrelation

Independence; (Moran's I ≈ -0.125)

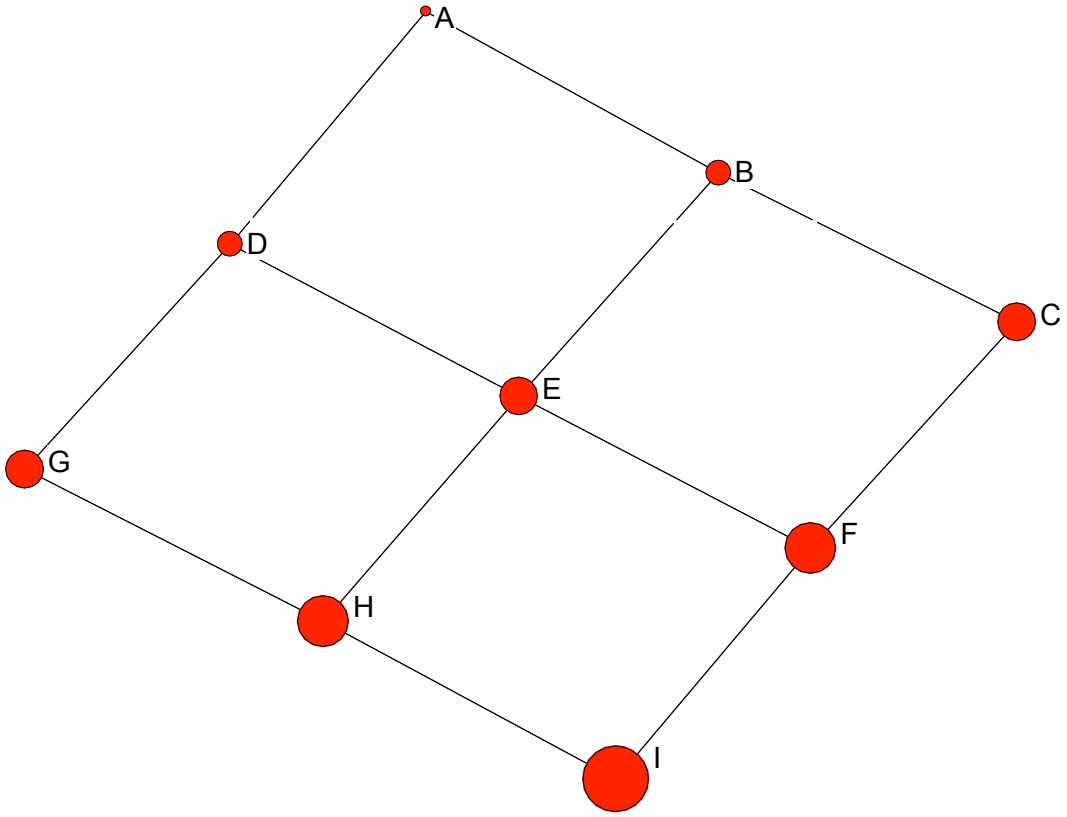



| Node | Attrib |
|---|---|
| A | 3 |
| B | 4 |
| C | 3 |
| D | 4 |
| E | 3 |
| F | 2 |
| G | 1 |
| H | 2 |
| I | 5 |

Moran's I: -0.250
Significance: 0.335

# Positive Autocorrelation

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

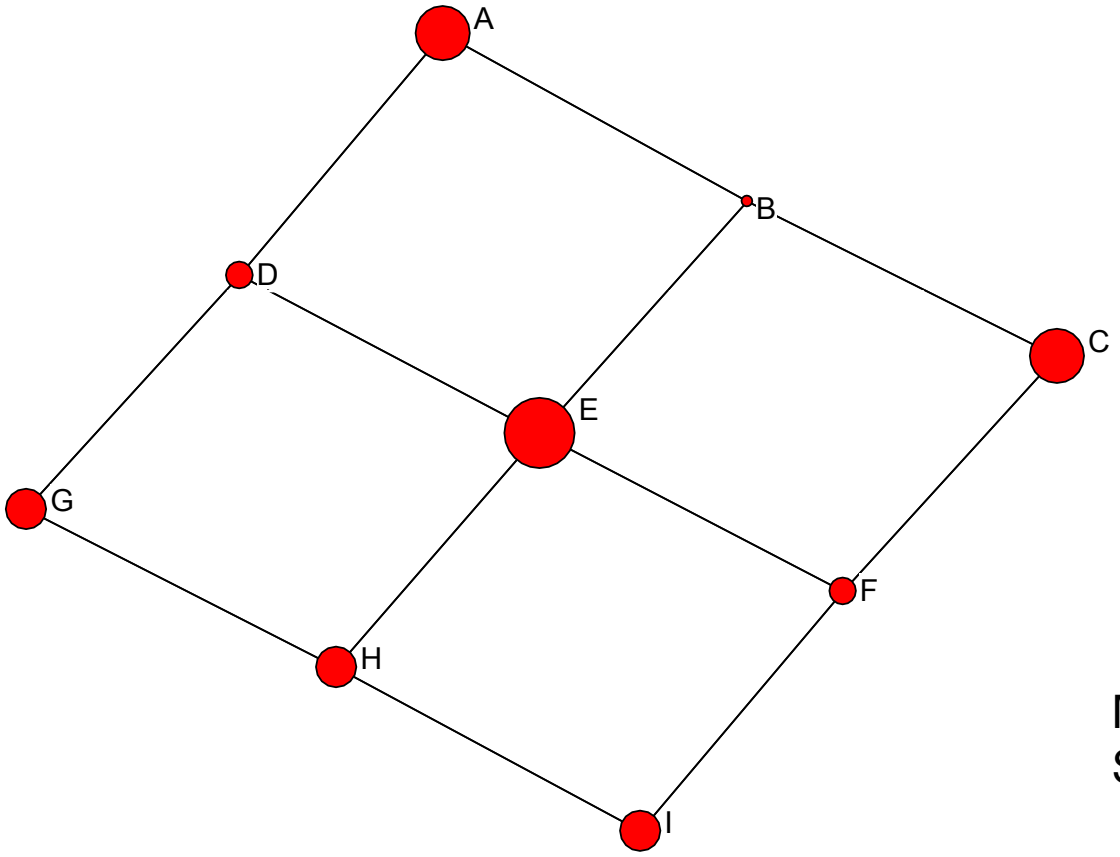



| Node | Attrib |
|---|---|
| A | 1 |
| B | 2 |
| C | 3 |
| D | 2 |
| E | 3 |
| F | 4 |
| G | 3 |
| H. | 4 |
| I. | 5 |

Moran's I: 0.500
Significance: 0.000

# 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

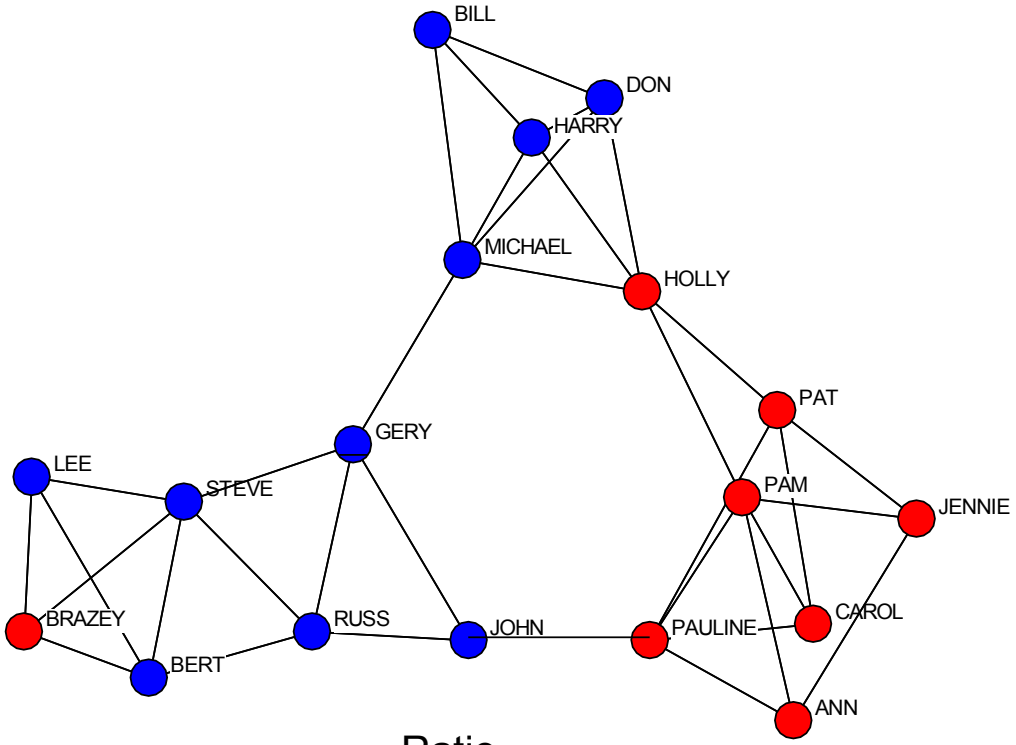
# Interpreting Autocorrelation

- 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 |

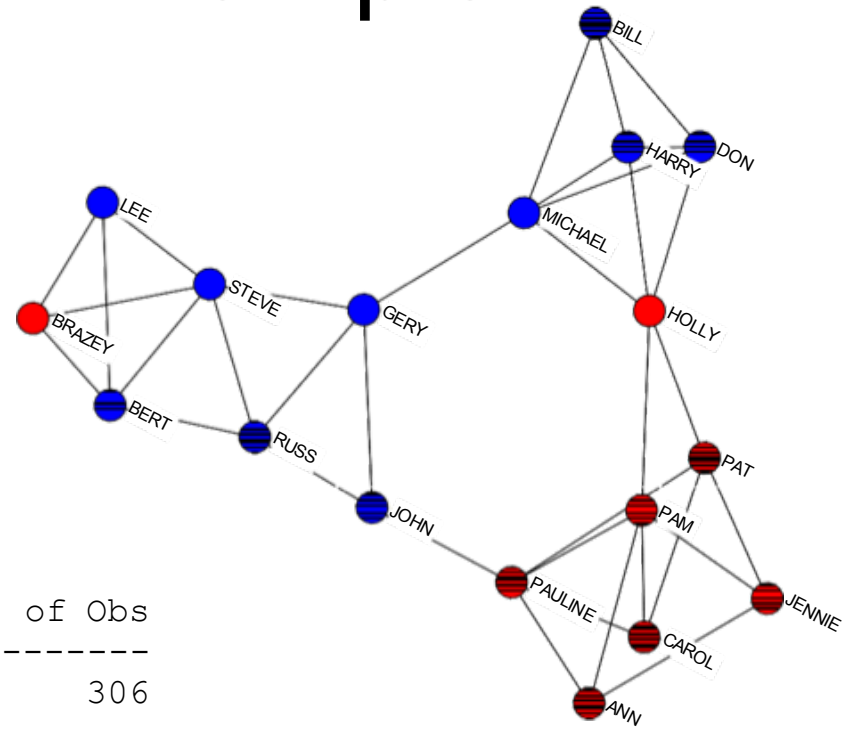
Expected

| | Female | Male |
|---|---|---|
| Female | 6.4 | 18.3 |
| Male | 18.3 | 10.3 |

# Campnet Example

Density Table

| | 1<br>Femal | 2<br>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

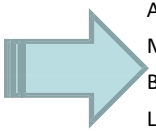
| Independent | Un-stdized Coefficient | Stdized Coefficient | Significance | Proportion As Large | Proportion As Small |
|---|---|---|---|---|---|
| Intercept | 0.087500 | 0.000000 | 1.000 | 1.000 | 0.001 |
| Group 1 | 0.341071 | 0.313982 | 0.001 | 0.001 | 0.999 |
| Group 2 | 0.268056 | 0.290782 | 0.001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

# 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