Social Network Analysis

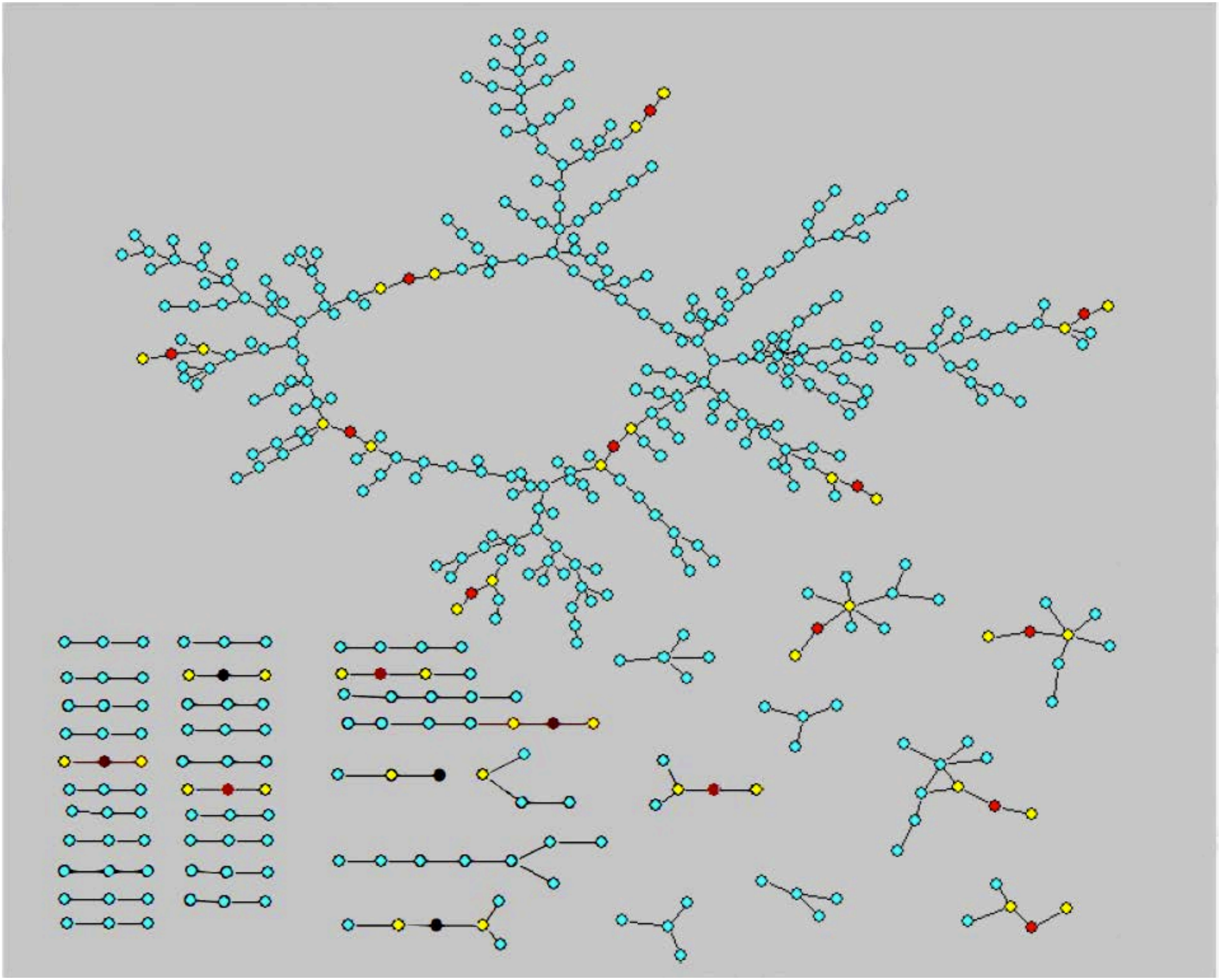
Paulo Serôdio

University of Essex

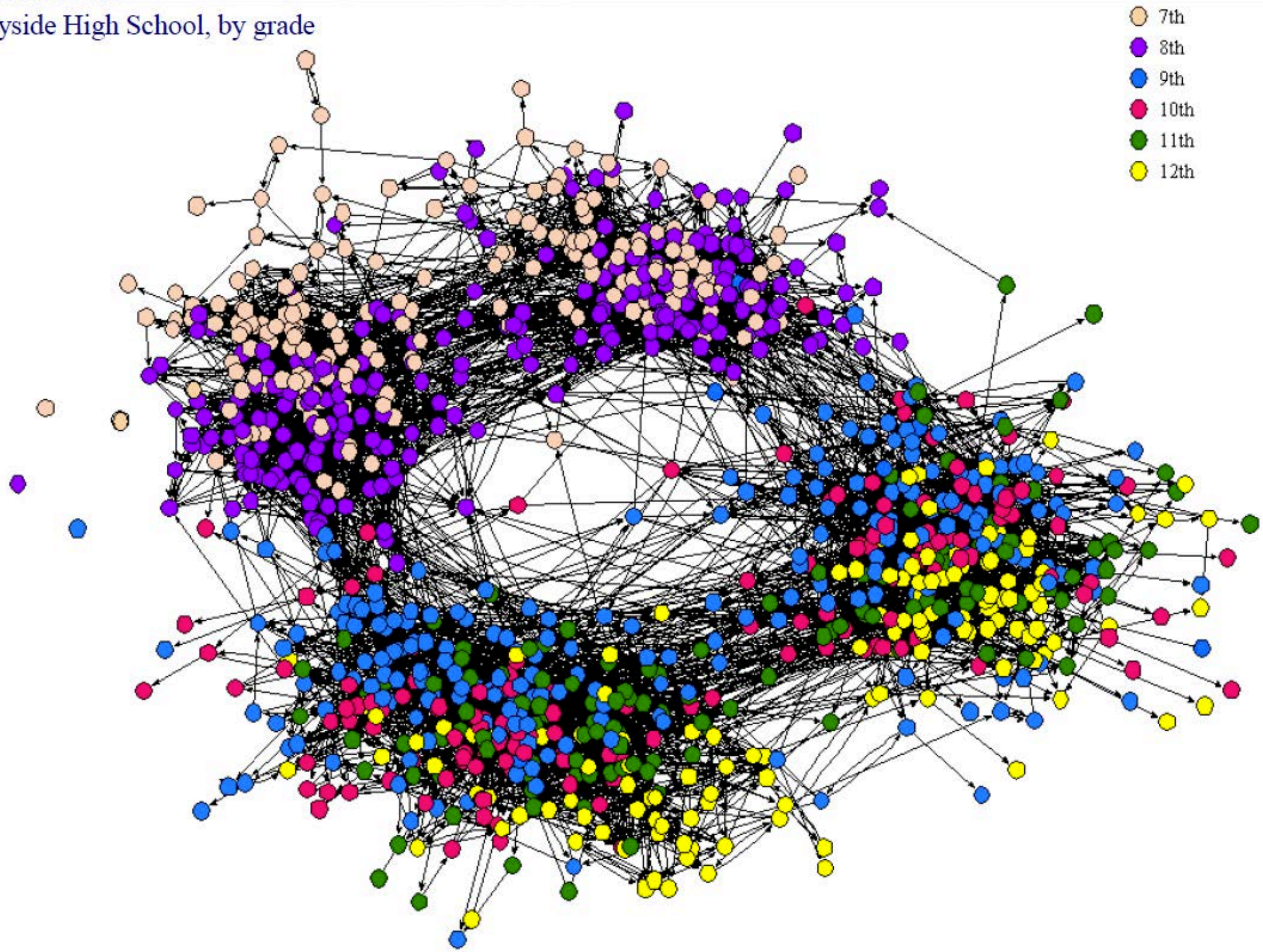
Part I - Data Collection

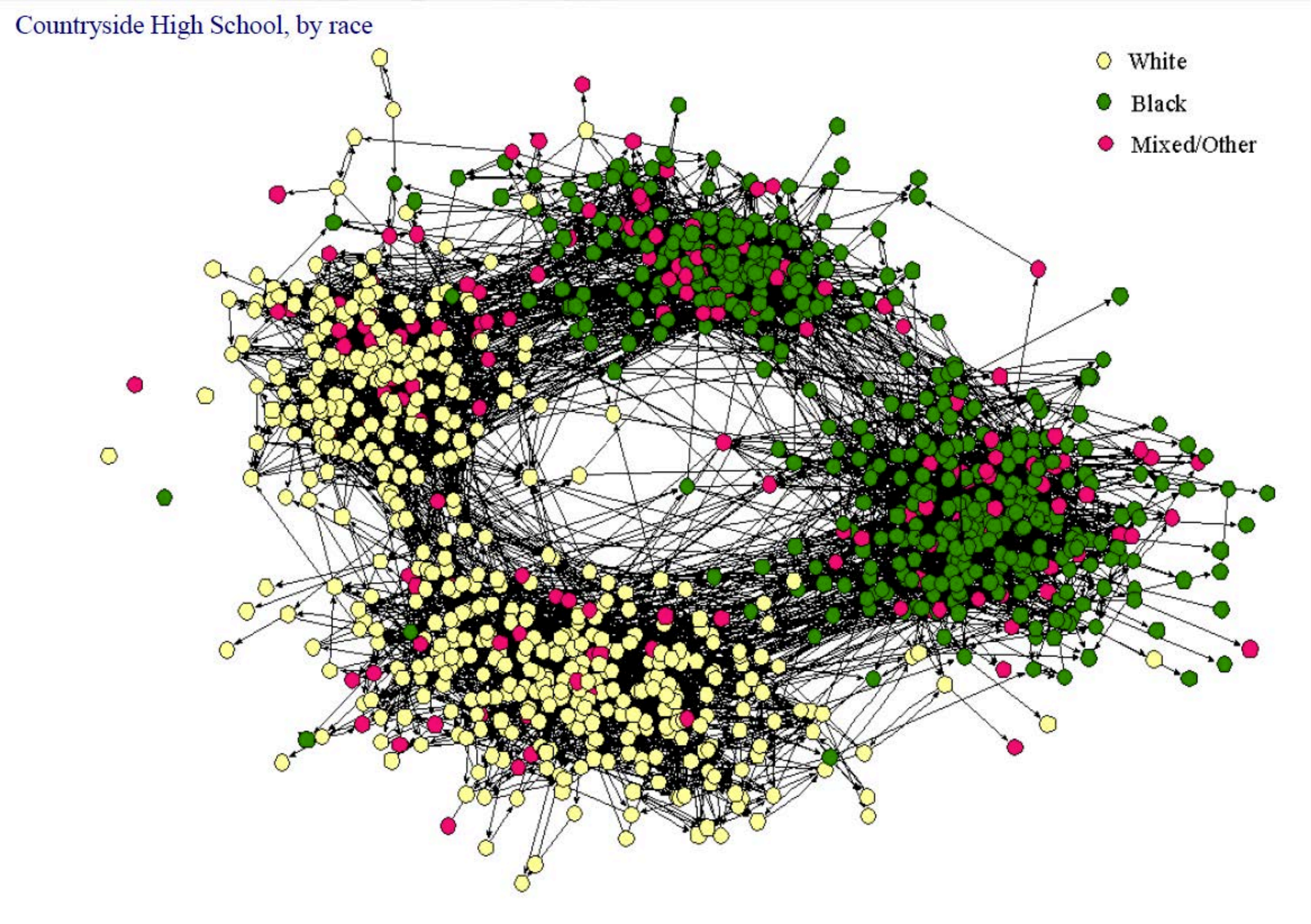
Structure Matters

- 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



Countryside High School, by grade




Countryside High School, by race
White
Black
Mixed/Other

Structure Matters

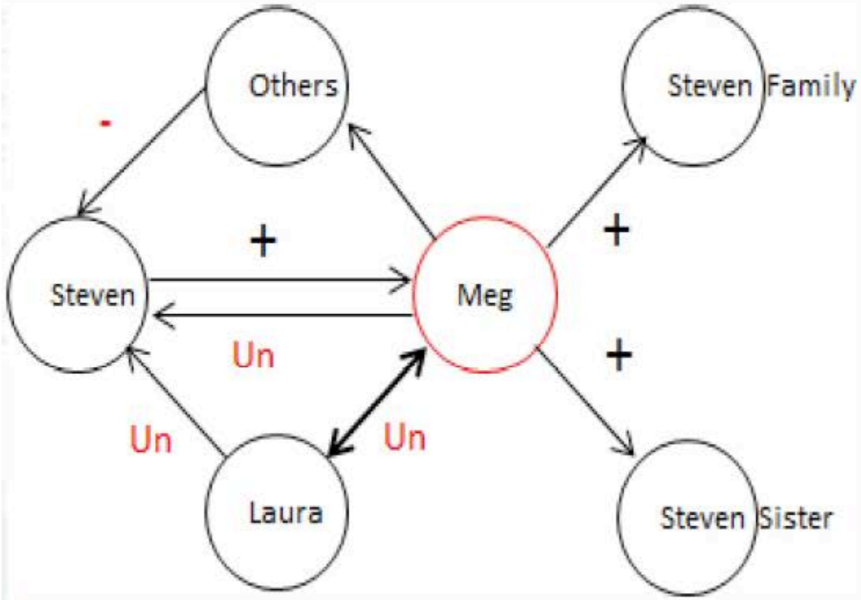
- SNA Core Research Goals
  - (1) Accurately represent social structures (descriptive)
    - Implications for outcomes (i.e. health)
  - (2) Explain how social structures come about, and what their consequences are (explanatory)
    - Ties forming and unforming
    - Actual measured outcomes (flows,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

- Network data is everywhere because social structure is everywhere!

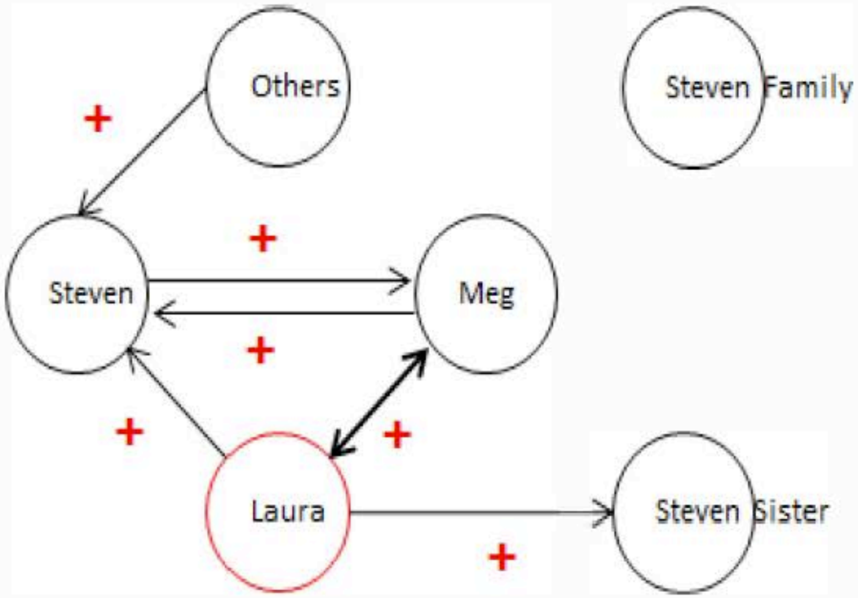
Meg … Do you know Steven Johanson? Alot of people think he’s a geek, I guess. But he likes me and he’s so nice. We talk on the phone alot and I went over to his house last night. Nothin’ happened but he is really nice and his family is nice, and he has a huge house and a pool. (Asshole! J/K) His sister is pretty, she doesn’t look 12 ½. She looks like she should be in 9th grade. A lot of people told me not to worry about what other people think. I asked him to TWIRP [“The Woman Is Required to Pay”-Dance] (kind of). I still have to figure out what’s happening. I don’t know what we’d do or where we’d go or who with. You’re probably thinking I’m crazy to go out with Steven, I hope you don’t think he’s a big nerd cuz I know he’s not super popular or anything, but not alot of people really know him, and once you get to know him, he’s super nice. Anyway, better go. W/B very soon.

Laura I know Steven pretty well, he’s a great guy. I think it would be awesome if you 2 went to TWIRP. He is just shy, not a big nerd, Sarah [his sister] is really pretty, we play tennis together.

Meg's View



Laura's View



Data Collection is Already Theory

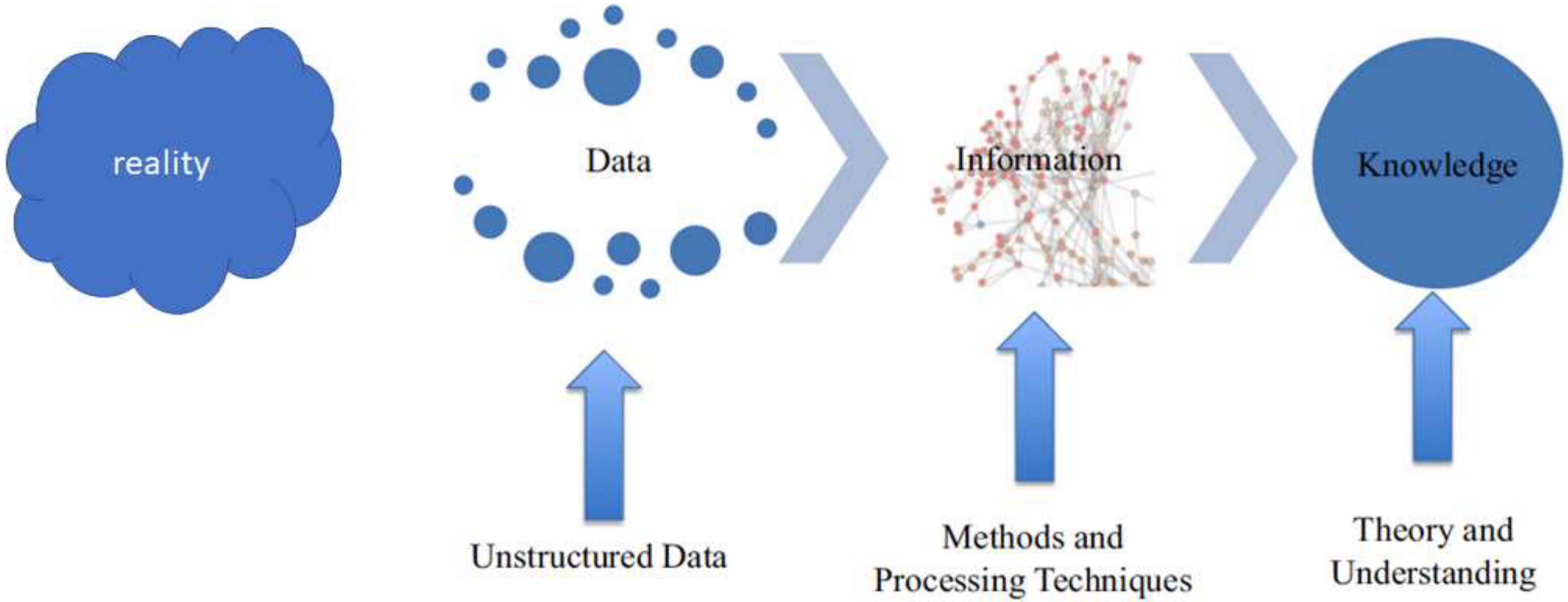


Fig. 1 Schematic of Data, Information, and Knowledge

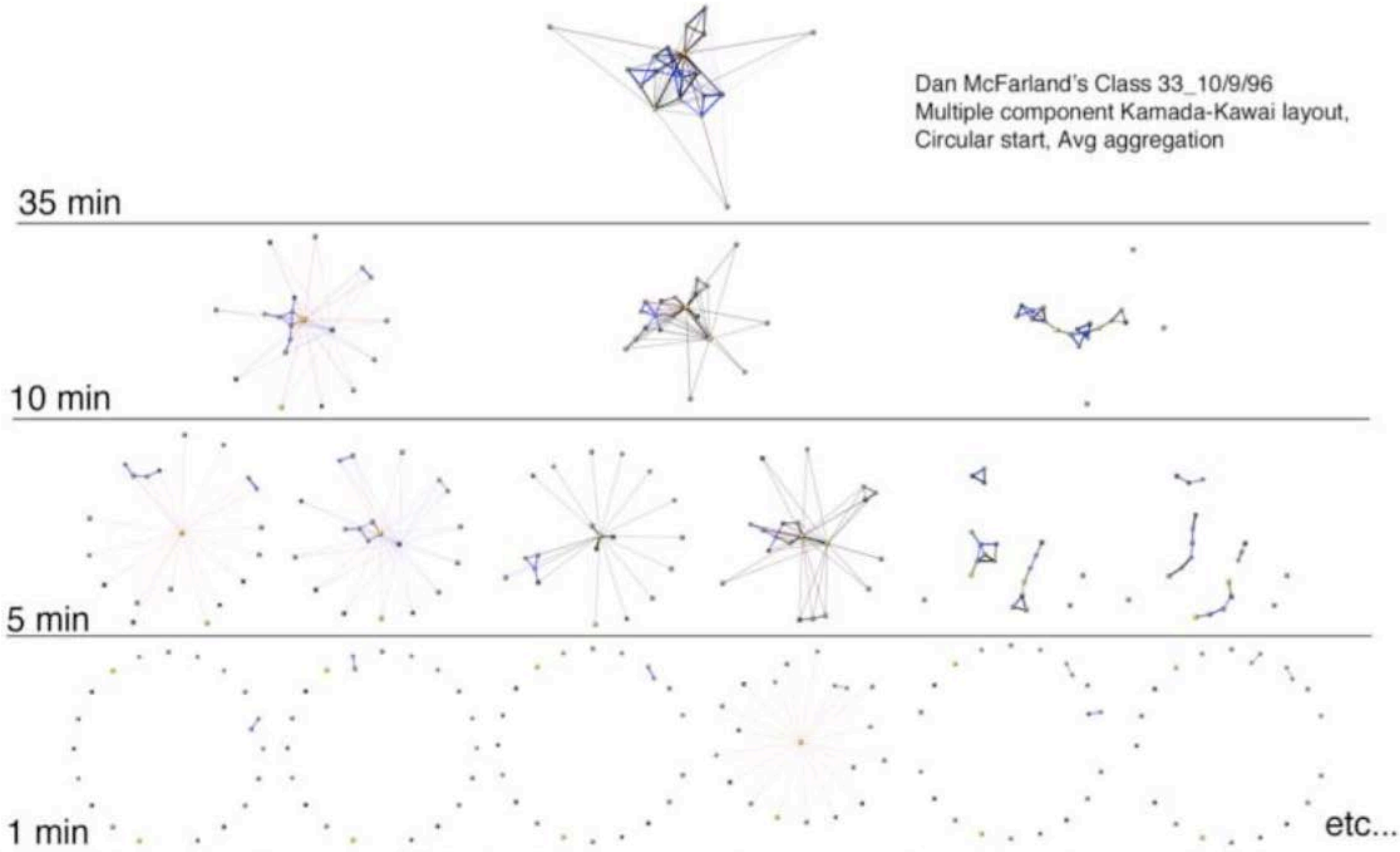


Figure 1. Interaction data from McFarland's classroom observations viewed at various levels of time aggregation from 35 minutes (one entire class period) to 1 minute (two to three turns of interaction).

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.

The Bank Wiring Room Study (Roethlisberger & Dickson, 1939)

•Investigate how **social dynamics** and **informal group norms** influence worker productivity.

Setup

• **Location**: Phone Banks Wiring Room at Western Electric's Hawthorne Works, Chicago.
• **Participants**: 14 male workers (9 wiremen, 3 soldermen, 2 inspectors) with interdependent tasks.
• **Duration**: 6 months of **non-intrusive observations**.

Data Collection Methods

• **Qualitative Observations**: Recorded interactions, communication patterns, and peer influence.
• **Productivity Records**: Tracked individual productivity to observe correlation with social dynamics.
• **Informal Social Network Analysis**: Documented friendships, alliances, and informal group norms.

Roethlisberger and Dickson 1939

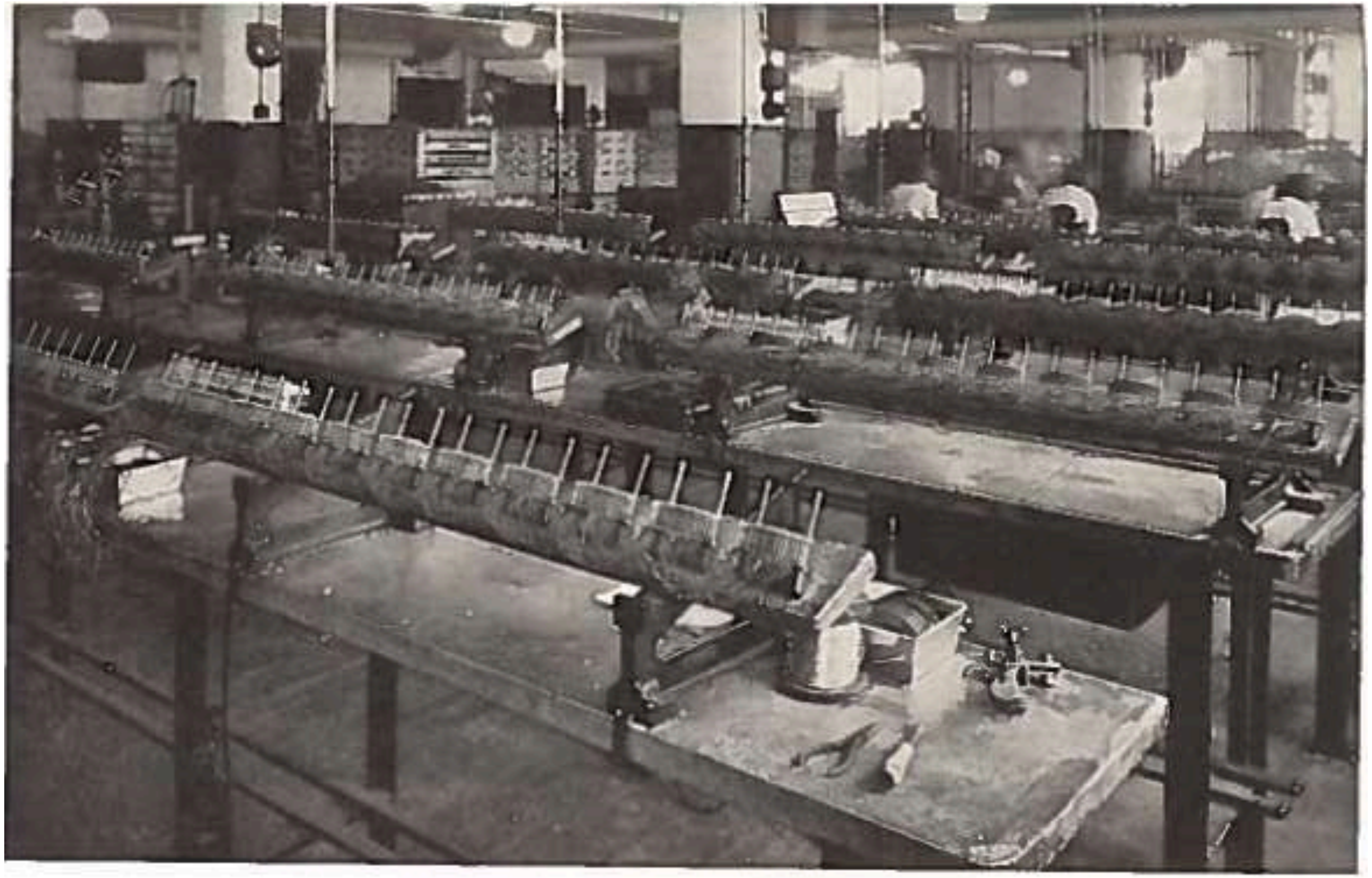


FIGURE 33
PHOTOGRAPH OF A SECTION OF THE BANK WIRING DEPARTMENT, SHOWING BANKS AT DIFFERENT STAGES OF COMPLETION

- 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: *Networks as pipes*	Diffusion Peer influence Social Capital "small worlds"	Social integration Peer selection Homophily Network robustness
Positional: *Networks as roles*	Popularity Effects Role Behavior Network Constraint	Group stability Network ecology "Structuration"

How Do Networks Form?

- Key Processes
  - Homophily
  - Shared Foci

  Exogenous Factors

  - Reciprocity
  - Transitive Closure
  - Preferential Attachment

  Endogenous Factors

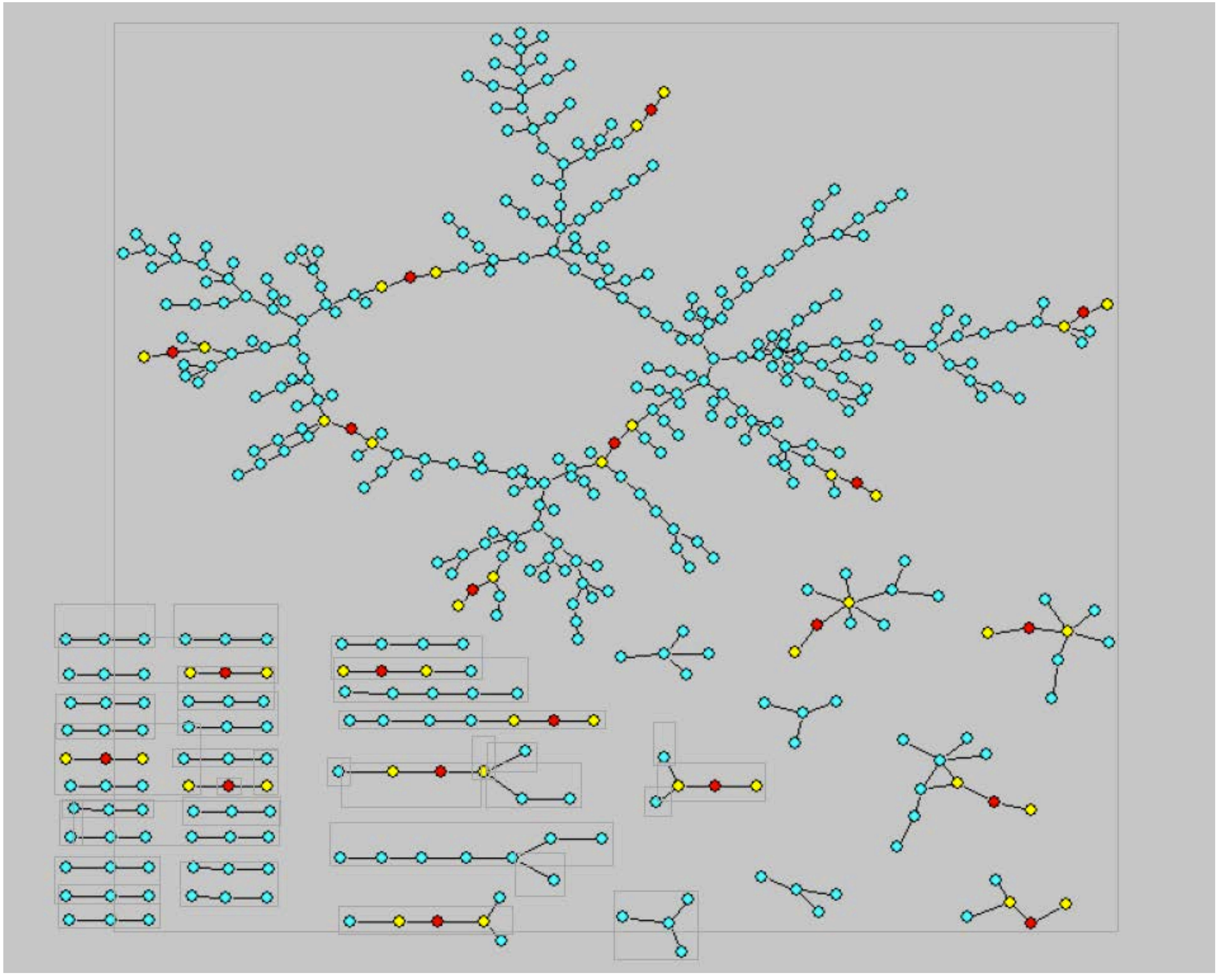
Defining Nodes & Ties

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

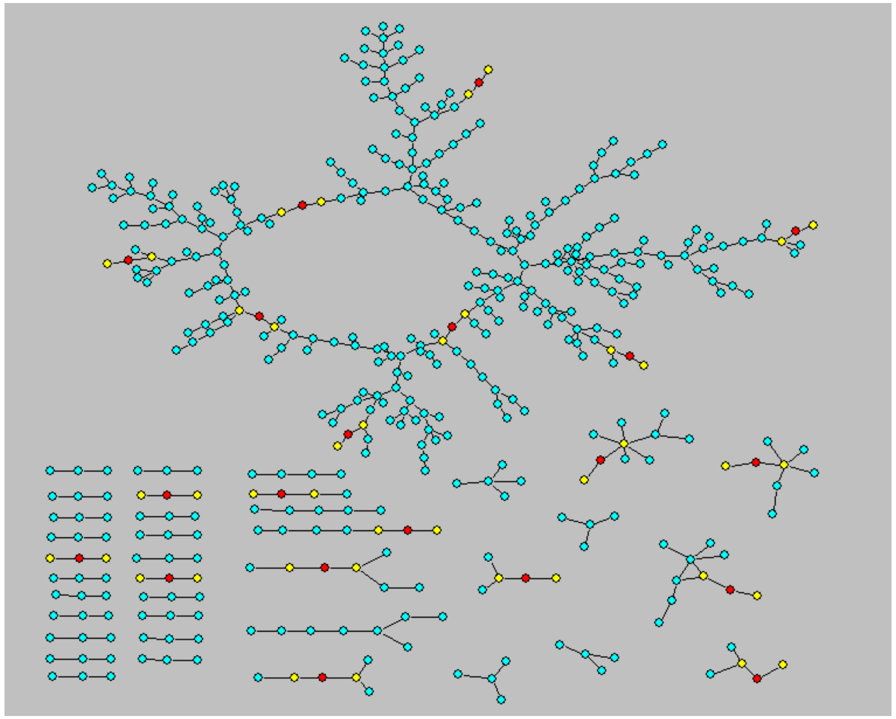
- **Levels of Analysis**
  - Individual ego, dyad, triad, clique/group/role, whole social structure

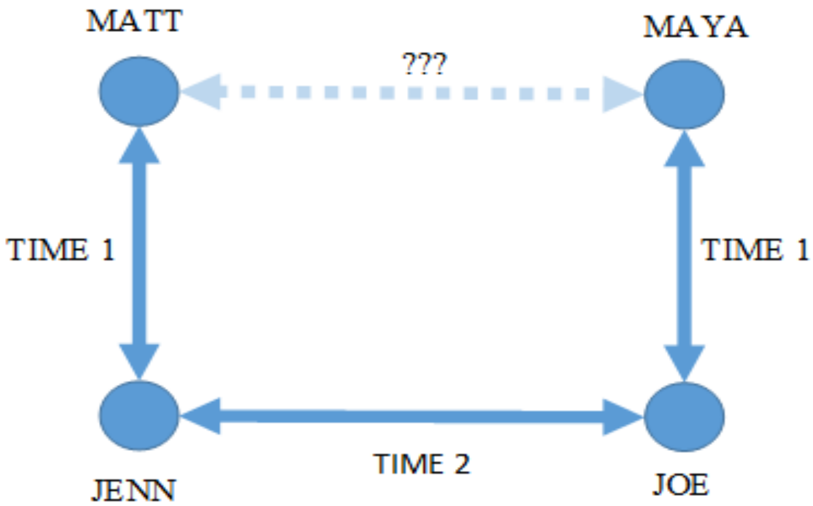
- **Units of time**
  - Seconds, minutes, hours, days, weeks, months, years, decades, centuries

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)

- **<u>Name Generator Surveys</u>**

  – **Free choice** (as many as you like) vs **Fixed choice** ("only top five")
    - Free >> Fixed choice: Issue of artificial cap – limited to <u>5</u> 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.

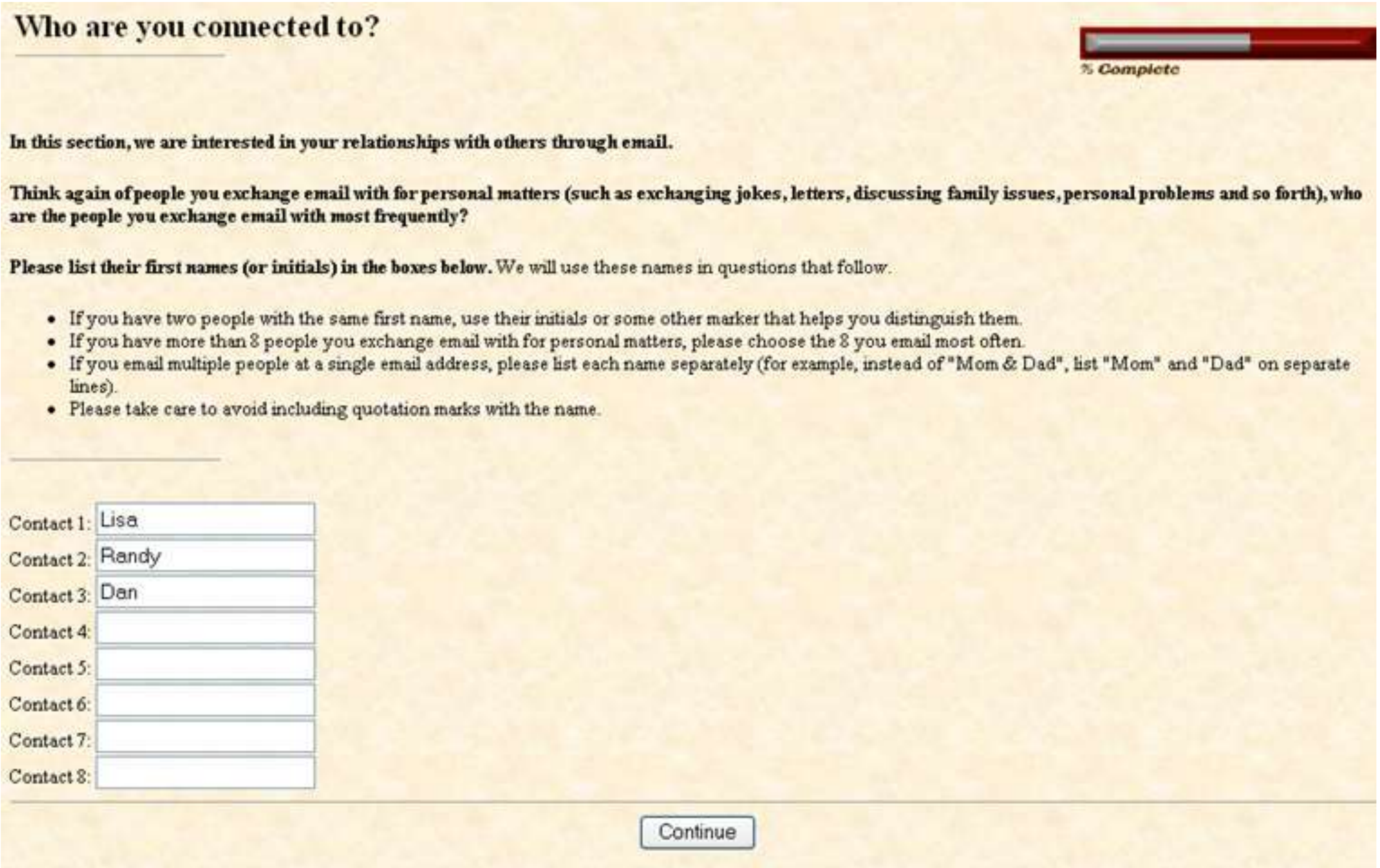
b) Question format:
- If you ask people to *recall* names (an open list format), fatigue will result in under-reporting
- If you ask people to check off names from a full list, you can often get over-reporting

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:



Who are you connected to?

% Complete

In this section, we are interested in your relationships with others through email.

Think again of people you exchange email with for personal matters (such as exchanging jokes, letters, discussing family issues, personal problems and so forth), who are the people you exchange email with most frequently?

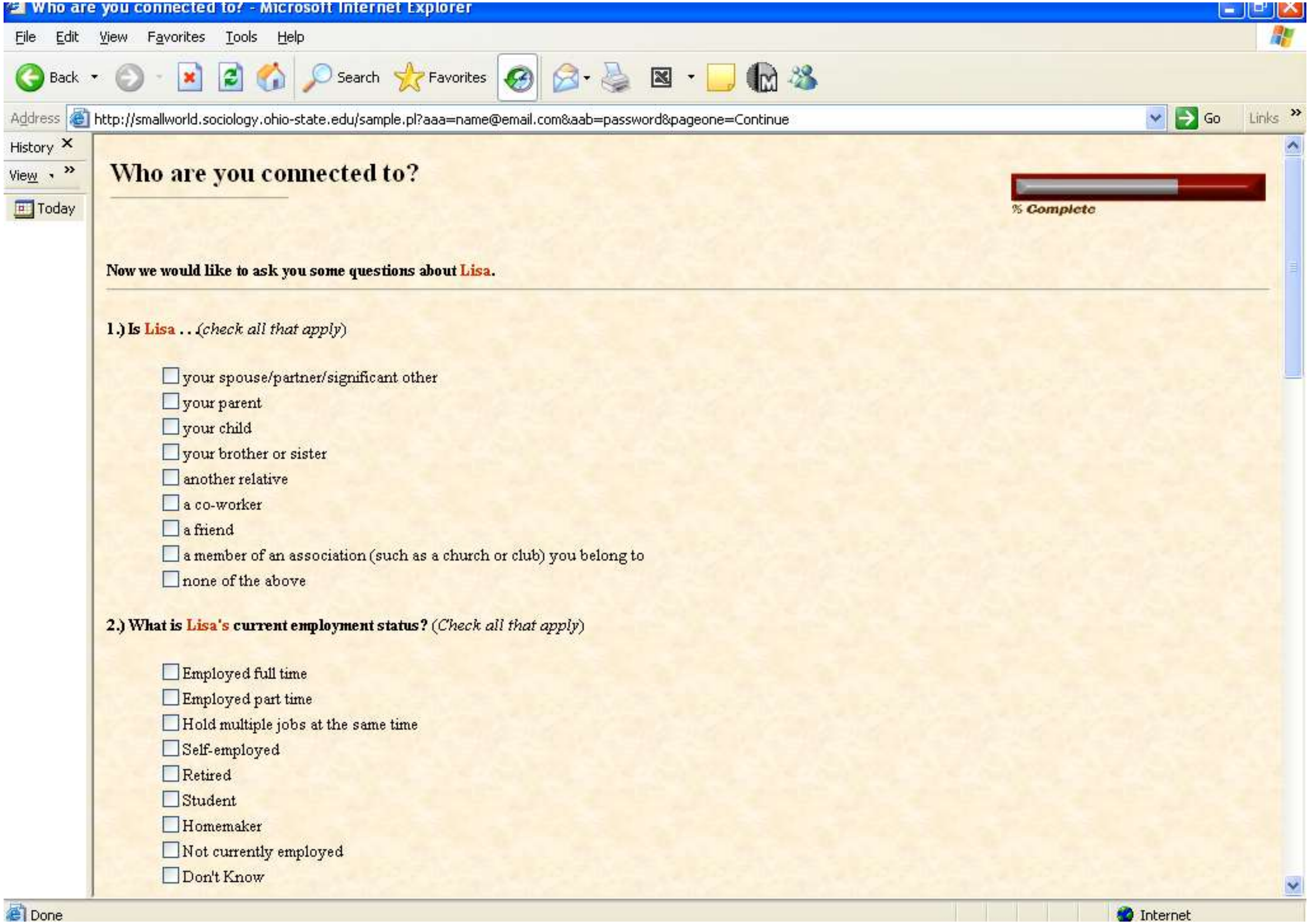
Please list their first names (or initials) in the boxes below. We will use these names in questions that follow.

- If you have two people with the same first name, use their initials or some other marker that helps you distinguish them.
- If you have more than 8 people you exchange email with for personal matters, please choose the 8 you email most often.
- If you email multiple people at a single email address, please list each name separately (for example, instead of "Mom & Dad", list "Mom" and "Dad" on separate lines).
- Please take care to avoid including quotation marks with the name.

Contact 1: Lisa
Contact 2: Randy
Contact 3: Dan
Contact 4:
Contact 5:
Contact 6:
Contact 7:
Contact 8:

Continue

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



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

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 _X_ Female	X	X	X						X	X	_X_ Yes ___ No	___ Yes _X_ No
(a)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(b)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(c)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(d)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(e)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(f)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(g)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(h)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(i)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No
(j)			___ Male ___ Female											___ Yes ___ No	___ Yes ___ No

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

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?" 1. Once a year or less 2. Every few months 3. Every few weeks 4. Once a week 5. Every day	"How often do you speak to each person on the list below?" 1. Very infrequently 2. Somewhat infrequently 3. About average 4. Somewhat frequently 5. Very frequently	1. Who do you talk to at least once every few months? (check all that apply) 2. Who do you talk to at least once every few weeks? 3. Who do you talk to at least once a week? 4. Who do you talk to every day?

• Need to do pre-testing to determine appropriate time scale
• Danger of getting no variance
• Assumes a lot from resps
	• Requires less of respondents; easier task
	• Is automatically normalized within respondent
	  • Removes response set issues
	  • Makes it hard to compare values across respondents (in different rows of data matrix)
		• Same data as absolute rating
		  • less tiring for respondent
		  • But questionnaire may look longer
		• With online surveys, can pipe responses so that respondent only sees names checked off in previous question
		• final question will have few names to react to

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

what question to ask?

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:

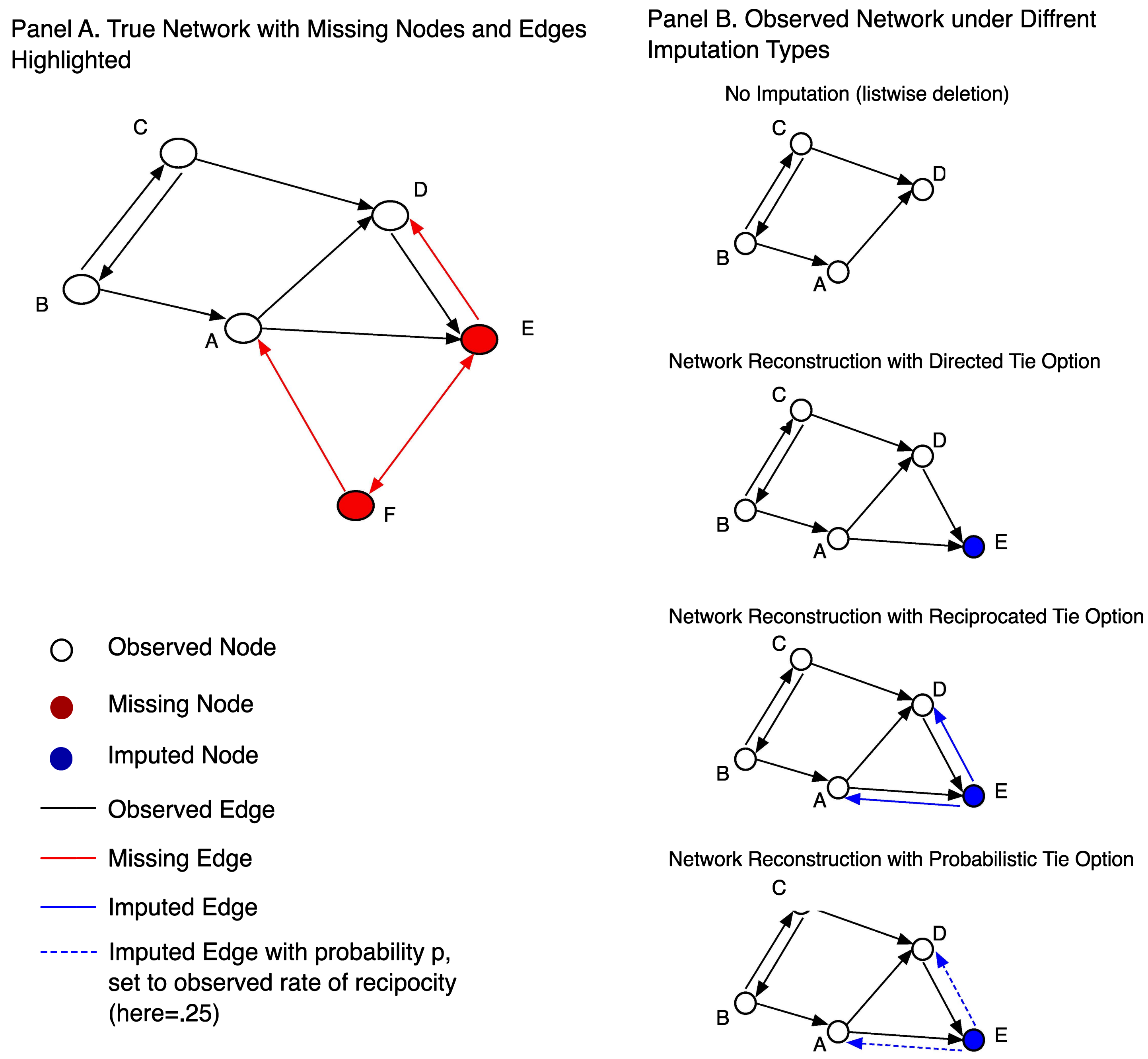
- Do nothing. If associated error is small ignore it. This is the default, not particularly satisfying.

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
C
D
B
A
E
F
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)
C
D
B
A
Network Reconstruction with Directed Tie Option
C
D
B
A
E
Network Reconstruction with Reciprocated Tie Option
C
D
B
A
E
Network Reconstruction with Probabilistic Tie Option
C
D
B
A
E

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?

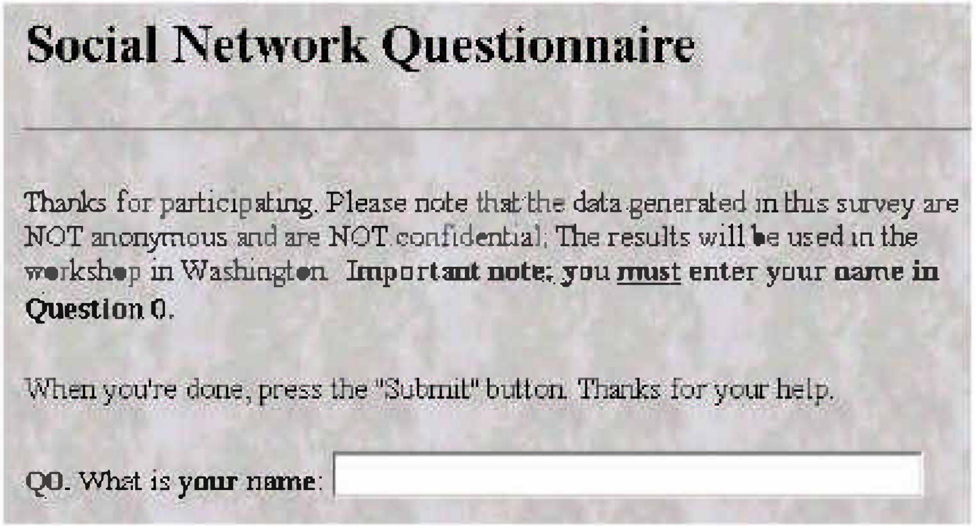
If a known-to-be STD positive person names a partner, do we inform the partner of the respondent's STD status?

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 <u>must</u> 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

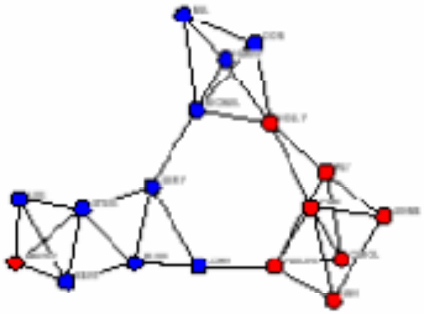
This is a social network study in which we will try to map out the communication network of the organization.

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

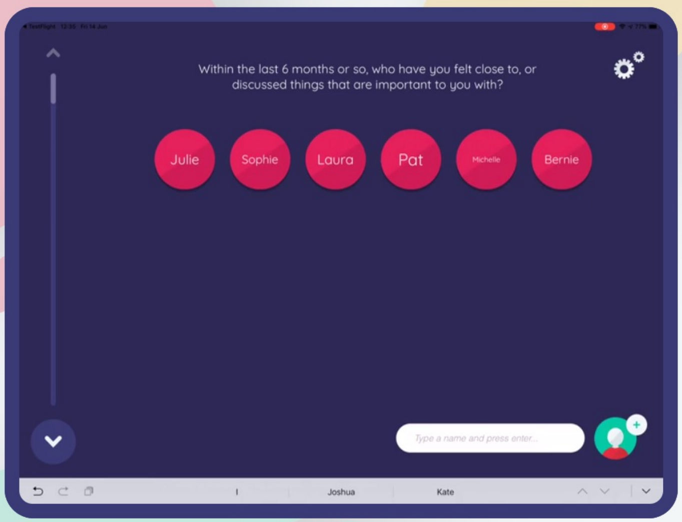
Who has access to identifying information

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

Network Canvas

COMMUNITY DOCUMENTATION PROJECTS **DOWNLOAD**

Simplifying complex network data collection.



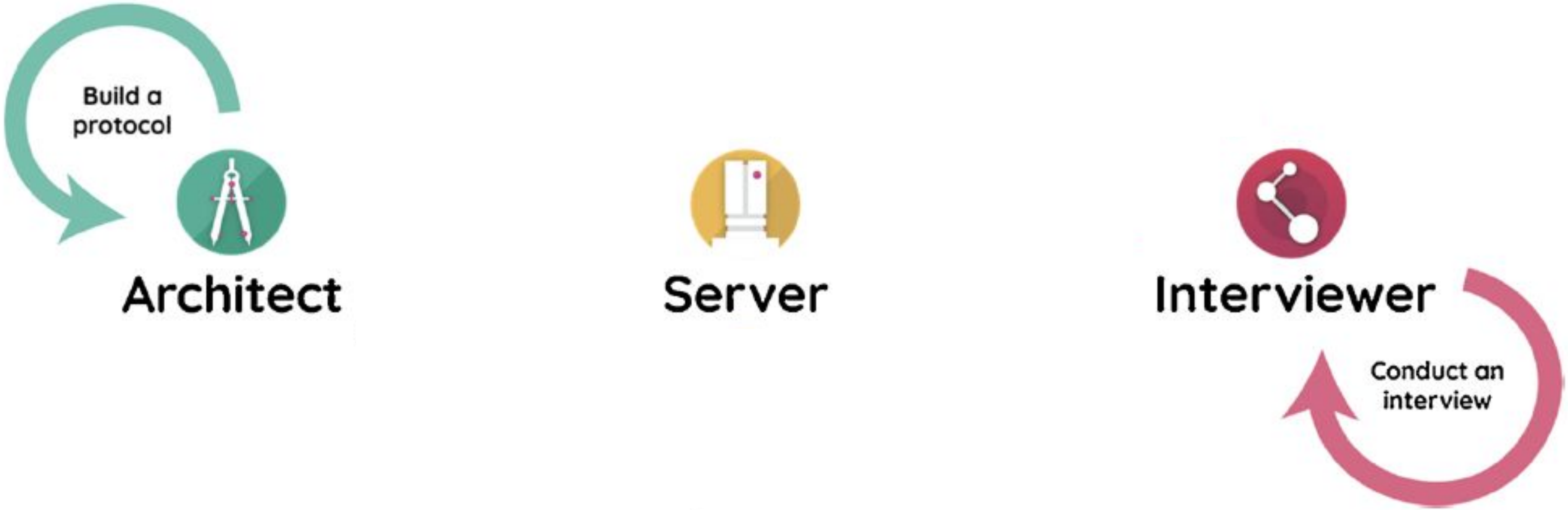
Network Canvas provides **free and open-source** software for surveying networks, designed around the needs of both researchers and their participants.

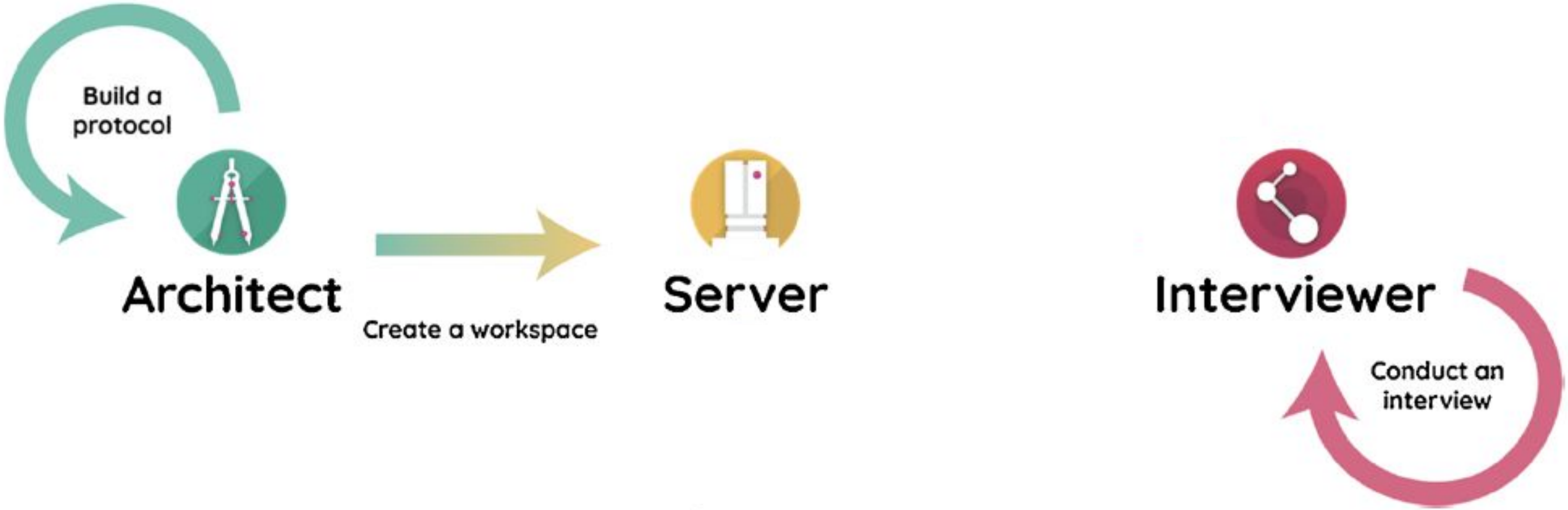
Architect

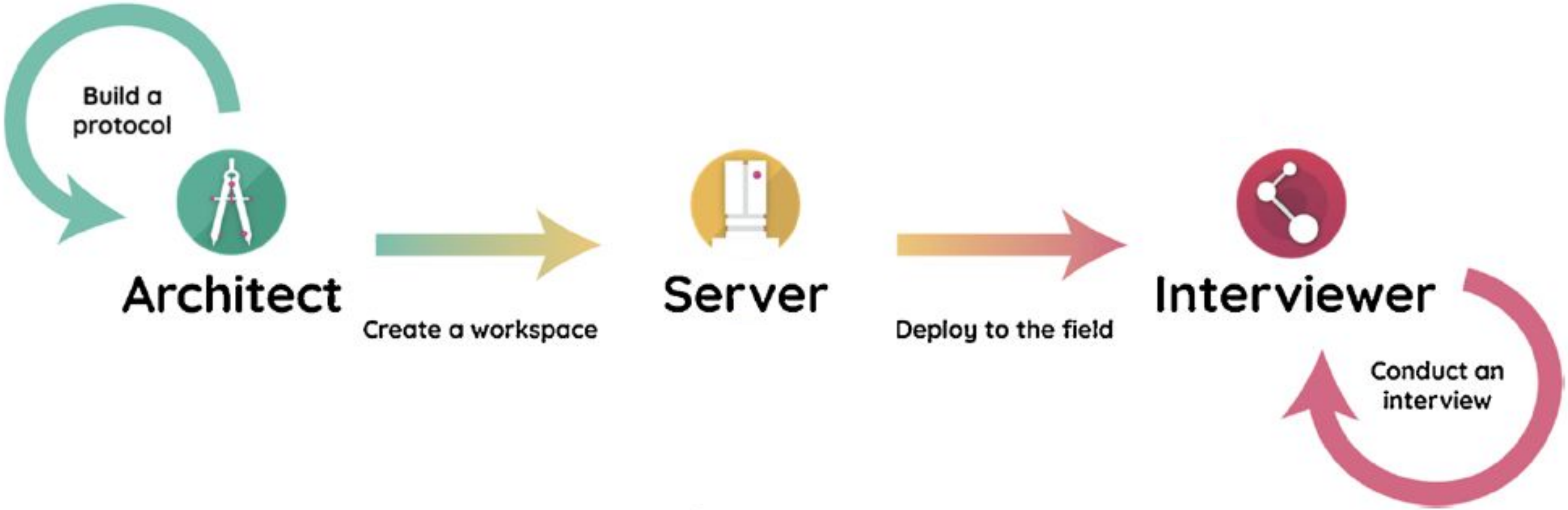
Server

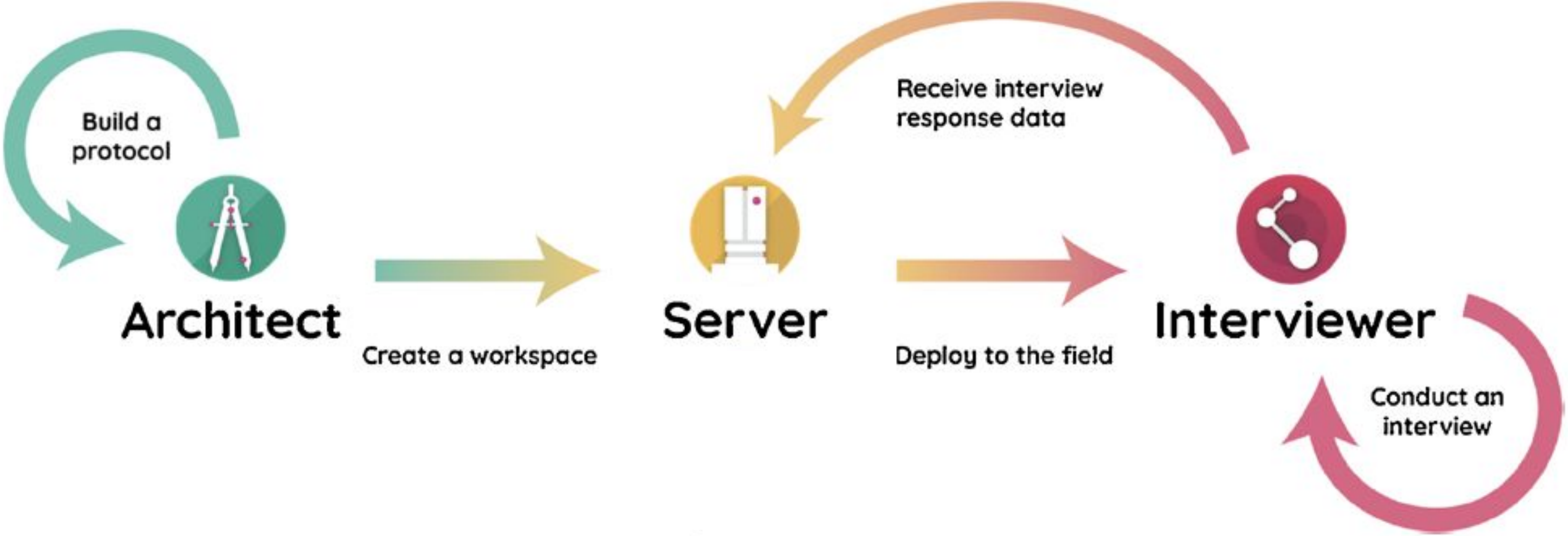
Interviewer

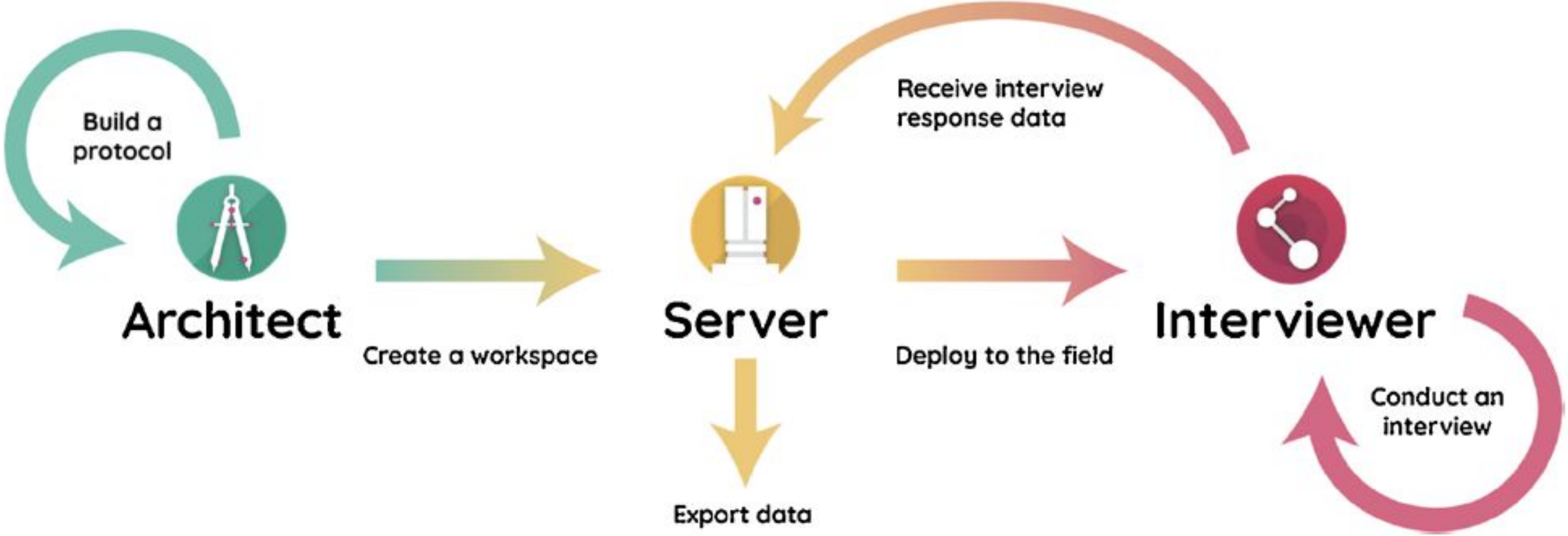

Build a protocol
Architect
Server
Interviewer


Build a
protocol
Architect
Server
Interviewer
Conduct an
interview


Build a protocol
Architect
Create a workspace
Server
Interviewer
Conduct an interview


Build a protocol
Architect
Create a workspace
Server
Deploy to the field
Interviewer
Conduct an interview


Build a protocol
Architect
Create a workspace
Server
Receive interview response data
Deploy to the field
Interviewer
Conduct an interview

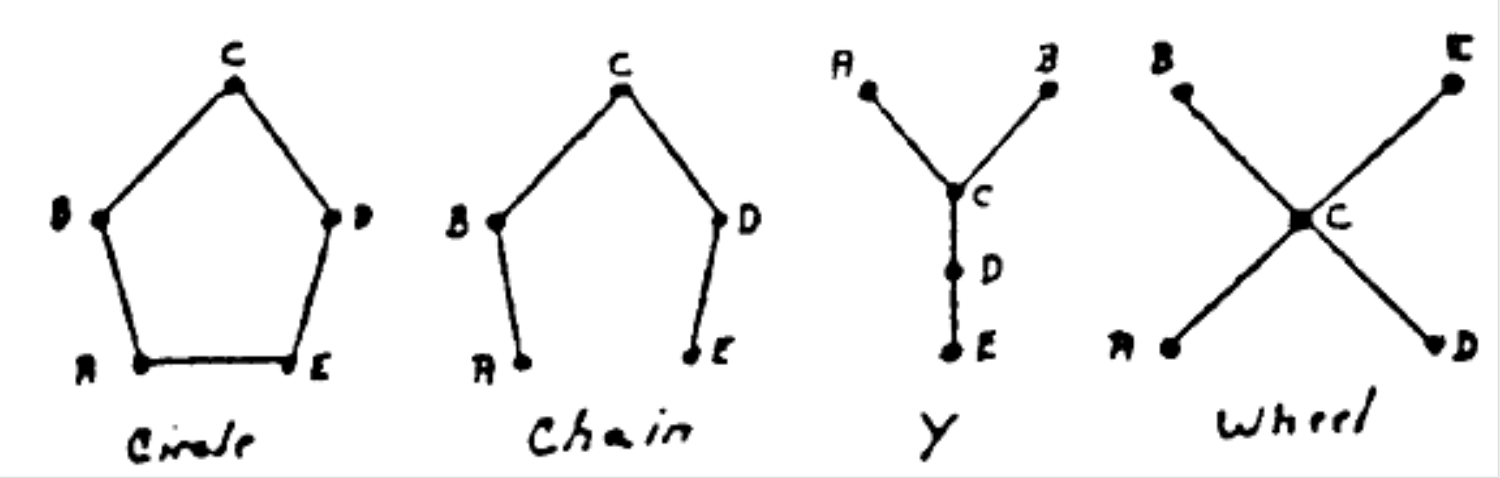

Build a protocol
Architect
Create a workspace
Server
Export data
Receive interview response data
Deploy to the field
Interviewer
Conduct an interview

Measures of Group Cohesion

<u>Whole Network Measures</u>

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

Bavelas-Leavitt experiments

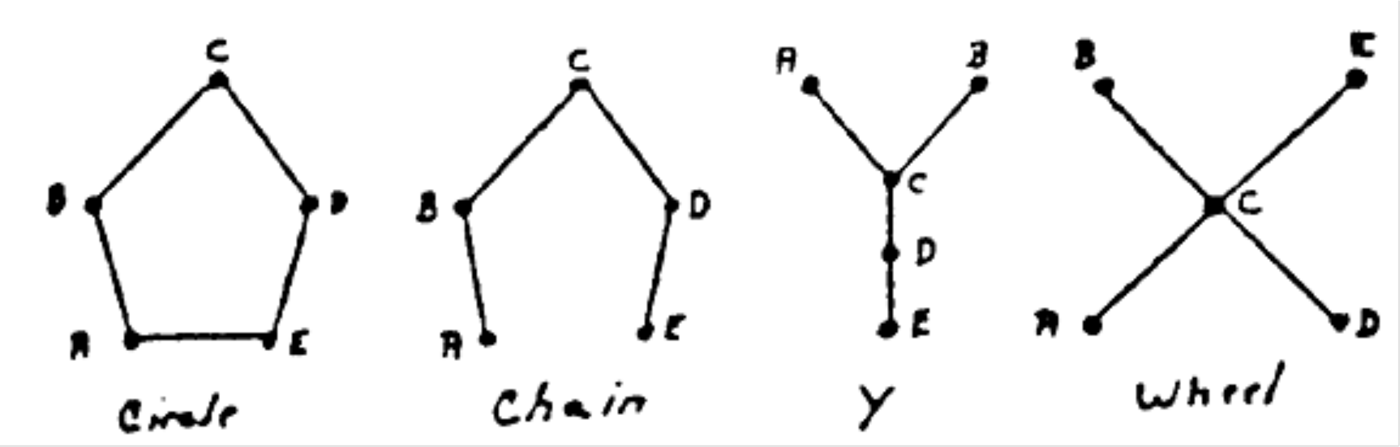


*Fastest possible time in units of number of moves

Each person can only send one message at a time.

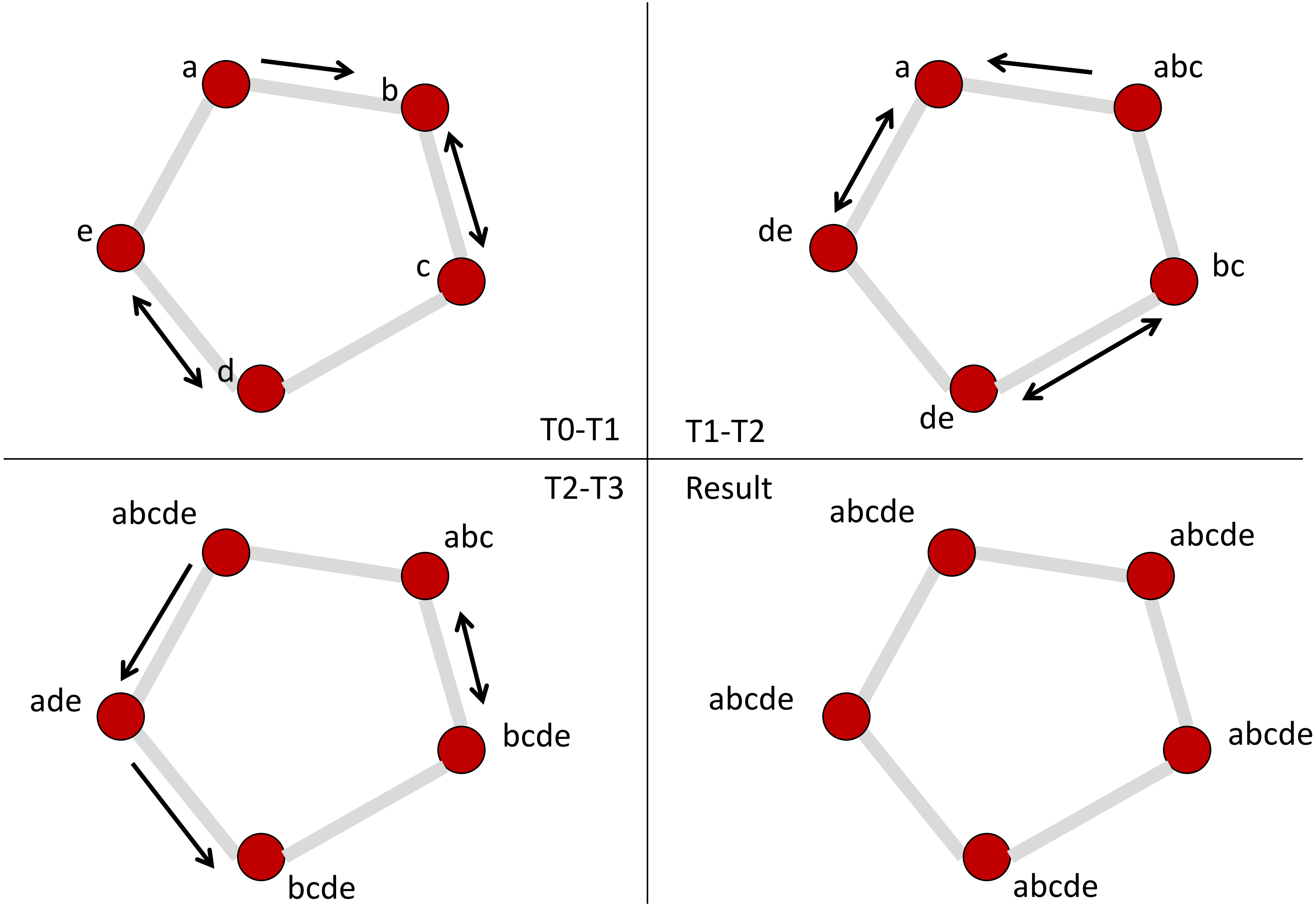
Bavelas-Leavitt experiments

Each person can only send one message at a time.



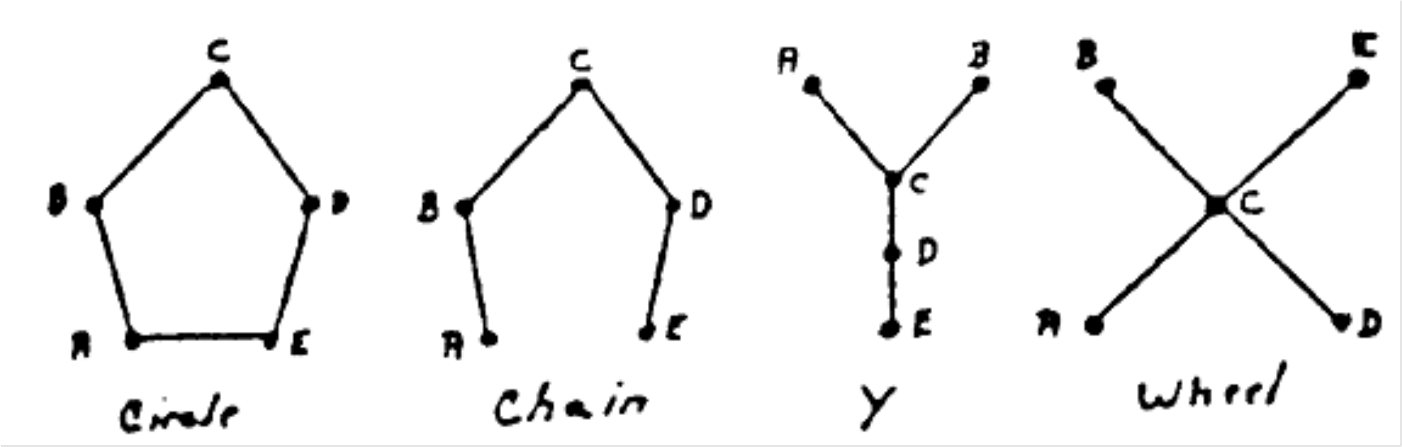
FPT*	3	5	4	5

*Fastest possible time in units of number of moves



Each person sends just one message aeceive multiple messages at one time.

Bavelas-Leavitt experiments



FPT	3	5	4	5
Time	50.4	53.2	35.4	32
No. of errors	7.6	2.8	0	0.6
No. of msgs	high	low	low	low

Key Findings

Expectation: Decentralized networks (e.g., Circle) should solve tasks faster, as information can flow freely without a central bottleneck.

Centralized Networks (Wheel, Y): Faster, fewer messages, fewer errors, clear leader identification.

Decentralized Networks (Circle): Higher satisfaction, more flexible but more prone to errors and inefficiency.

Why?

• **Centralization Effect**: In centralized structures, information funnels to a central "integrator" (clear leader), making it easier for participants to follow a single, efficient strategy without confusion.

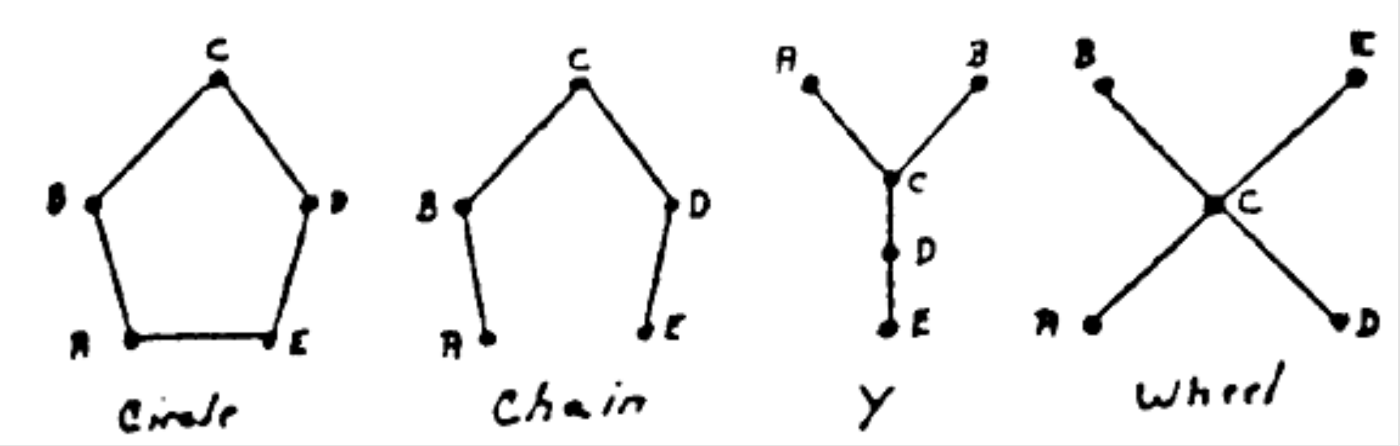
• **Complexity of Decentralized Systems**: Decentralized networks, while theoretically efficient, offer many possible communication paths, creating choice overload and coordination issues. This lack of a forced strategy made it harder for participants to align and solve tasks quickly.

• **Cognitive Preference for Leadership**: Participants naturally gravitate toward clear, hierarchical structures (centralized systems) where leadership and roles are obvious, making problem-solving more intuitive even if it's not mathematically optimal.

Bavelas-Leavitt interpretation

- In centralized networks, the distance from the “natural integrator”
- Centralization is good for simple, routine tasks

Total distance from “natural integrator”			
6	6	5	4



Measuring Bavelas centralization

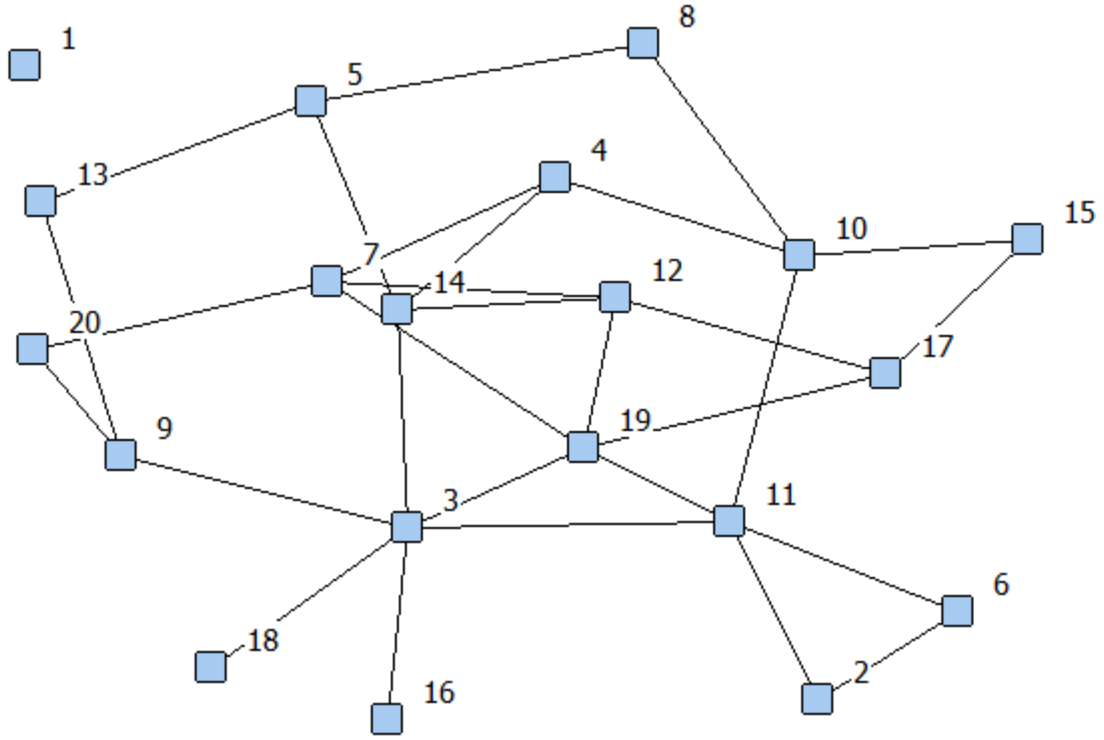
- Calculate graph-theoretic distances between every node and every other
- Find the node least far from all the others (e.g., smallest avg dist)
  - Call this the center
- Sum the the distances of every node to the center
  - This is Bavelas centralization
- See also Freeman’s closeness centralization

Characterizing whole networks

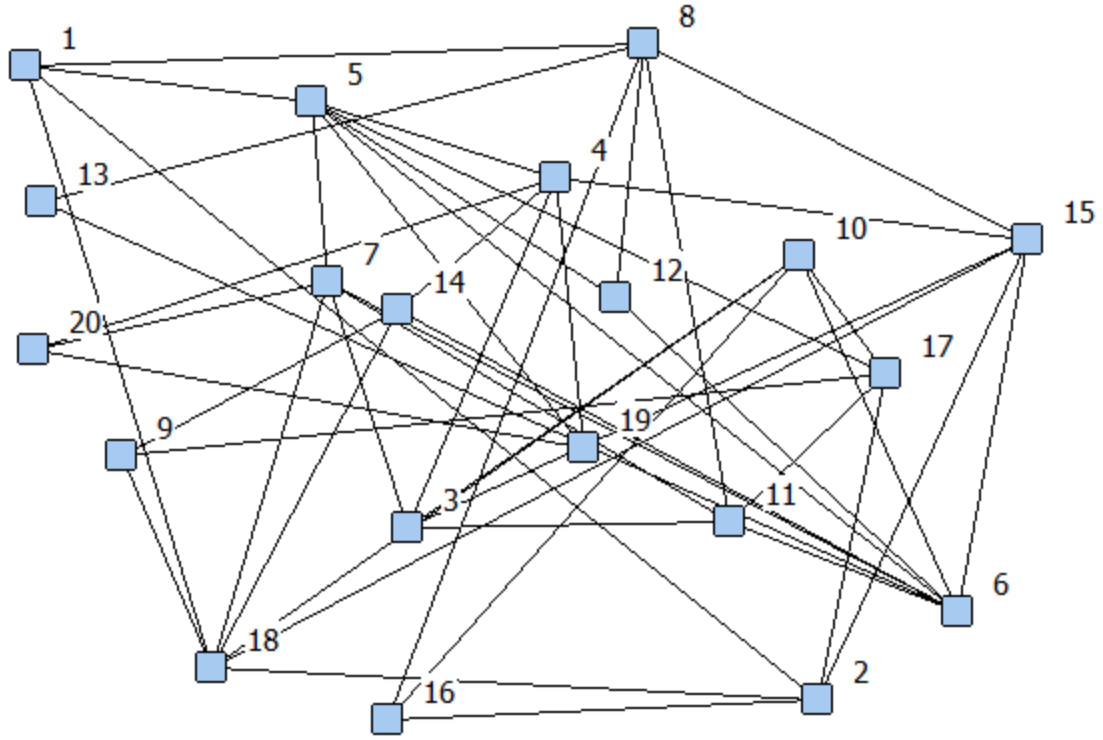
- Cohesion is biggest topic
  - Most measures of cohesion come from summarizing lower-level indices
    - E.g. average tie strength (aka density)
- There are also measures of shape
  - Many of these are "configural" in the sense that they are not simple aggregations of lower-level measures
    - E.g., core-periphery measures

Density

- Number of ties, expressed as proportion of # possible



Density = 0.15



Density = 0.25

Density

- Density is the number of ties in the network as a whole, expressed as proportion of # possible

	Reflexive	Non-Reflexive
Undirected	$= \frac{T}{n^2/2}$	$= \frac{T}{n(n-1)/2}$
Directed	$= \frac{T}{n^2}$	$= \frac{T}{n(n-1)}$

T = number of ties in network
n = number of nodes

Density as aggregated dyadic cohesion (or normalized node degree)

	HOLLY	BILL	DON	HARRY	MICHAEL	PAM	JENNIE	ANN	PAULINE	PAT	CAROL	LEE	JOHN	BRAZEY	GERY	STEVE	BERT	RUSS	Avg
HOLLY		0	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0.294
BILL	0		1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.176
DON	1	1		1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.235
HARRY	1	1	1		1	0	0	0	0	0	0	0	0	0	0	0	0	0	0.235
MICHAEL	1	1	1	1		0	0	0	0	0	0	0	0	0	1	0	0	0	0.294
PAM	1	0	0	0	0		1	1	1	0	1	0	0	0	0	0	0	0	0.294
JENNIE	0	0	0	0	0	1		1	0	1	0	0	0	0	0	0	0	0	0.176
ANN	0	0	0	0	0	1	1		1	0	0	0	0	0	0	0	0	0	0.176
PAULINE	0	0	0	0	0	1	0	1		1	1	0	1	0	0	0	0	0	0.294
PAT	1	0	0	0	0	0	1	0	1		1	0	0	0	0	0	0	0	0.235
CAROL	0	0	0	0	0	1	0	0	1	1		0	0	0	0	0	0	0	0.176
LEE	0	0	0	0	0	0	0	0	0	0	0		0	1	0	1	1	0	0.176
JOHN	0	0	0	0	0	0	0	0	1	0	0	0		0	1	0	0	1	0.176
BRAZEY	0	0	0	0	0	0	0	0	0	0	0	1	0		0	1	1	0	0.176
GERY	0	0	0	0	1	0	0	0	0	0	0	0	1	0		1	0	1	0.235
STEVE	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1		1	1	0.294
BERT	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1		1	0.235
RUSS	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1		0.235
Avg	0.29	0.18	0.24	0.24	0.29	0.29	0.18	0.18	0.29	0.24	0.18	0.18	0.18	0.18	0.24	0.29	0.24	0.24	0.229

Density tables

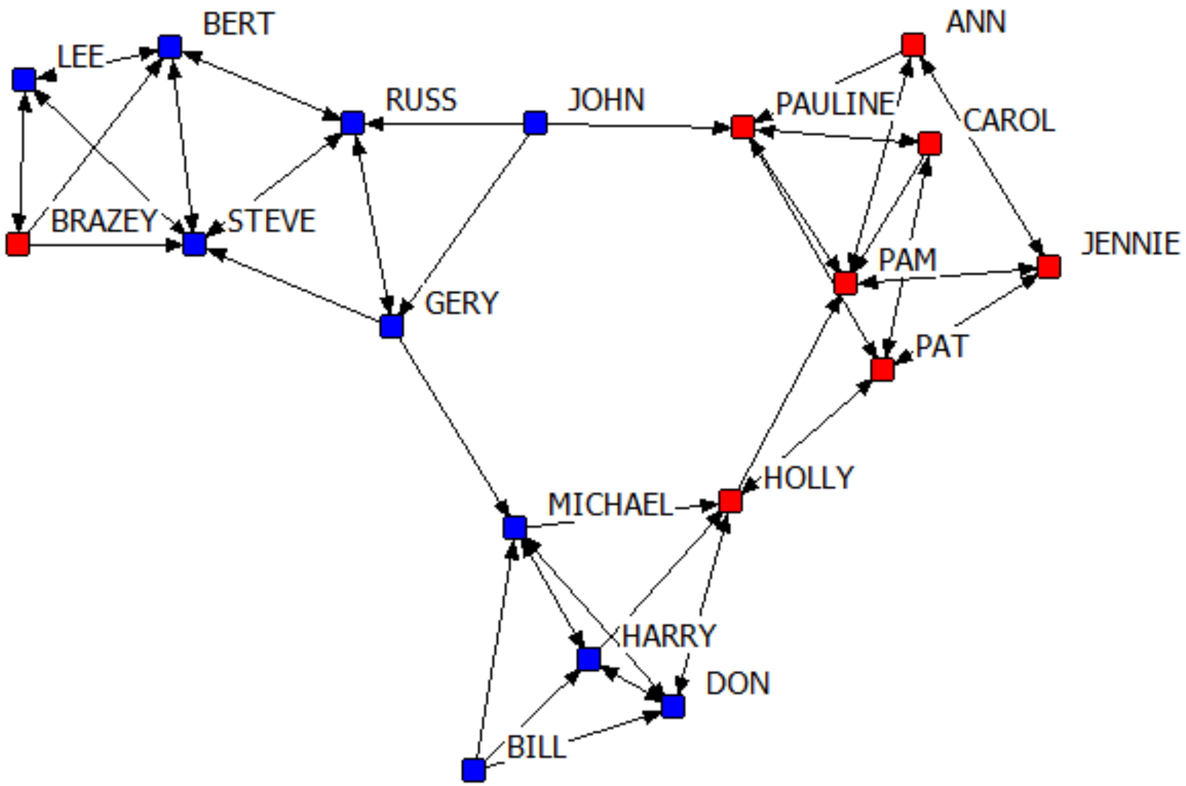
- Density of ties within and between *a priori* groups

Number of ties

	1	2
1	20	4
2	5	25

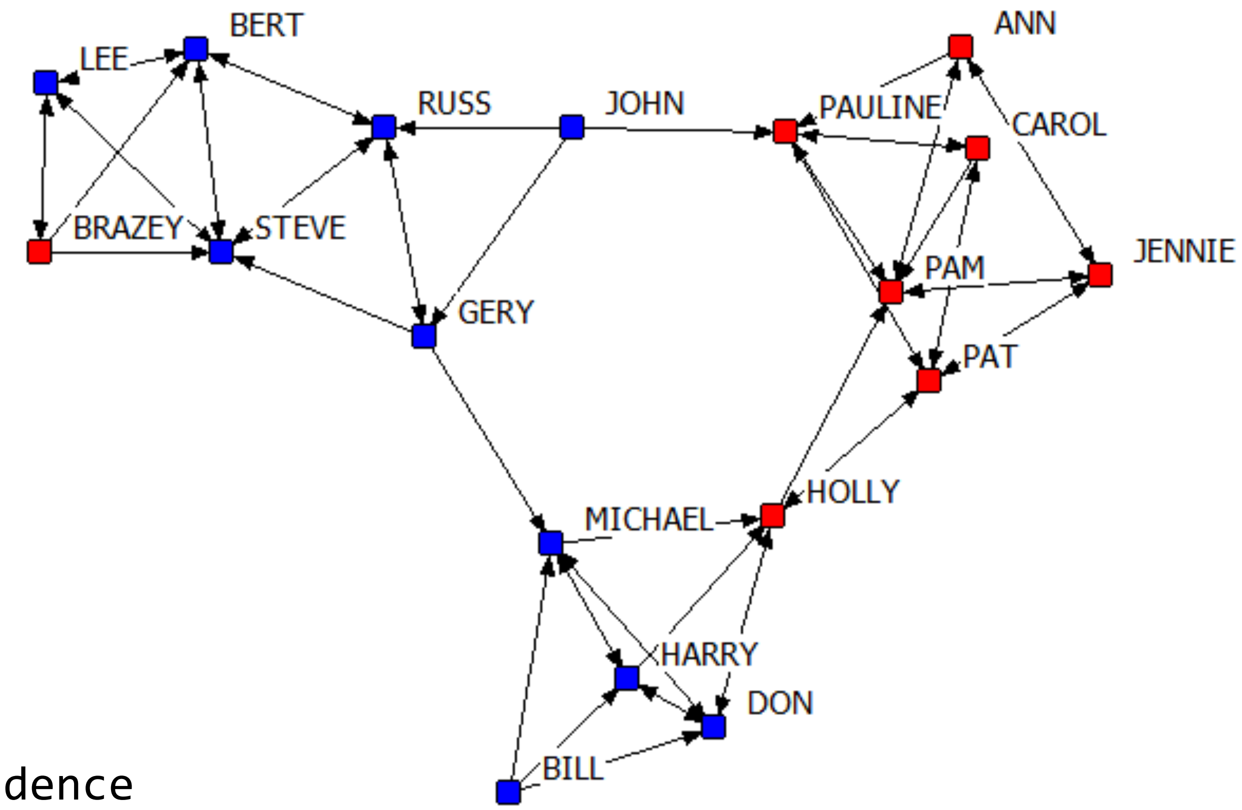
Density of ties

	1	2
1	0.357	0.050
2	0.063	0.278



Density tables

- Density of ties within and between *a priori* groups



Number of ties

	1	2
1	20	4
2	5	25

Expected Values Under Model of Independence

	1	2
1	9.88	14.12
2	14.12	15.88

```
Observed chi-square value = 28.732
Significance = 0.000100
```

“De-Energizing” Work Ties

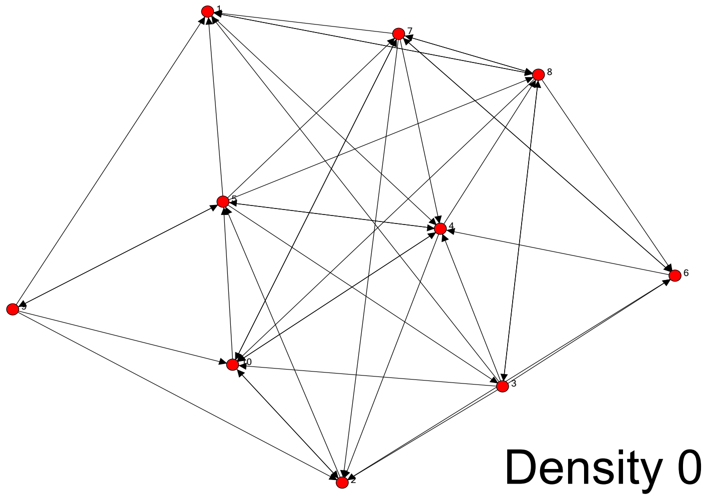
tie = "who tends to de-energize you?", run at a pizza supplier, symmetrized.

- Cross department Interactions
- 36 dept-to-dept work interaction pairs
- 7 pairs have >= 10% de-energizing work interactions
- Departments #6 and #9 have 50% de-energizing interactions between them

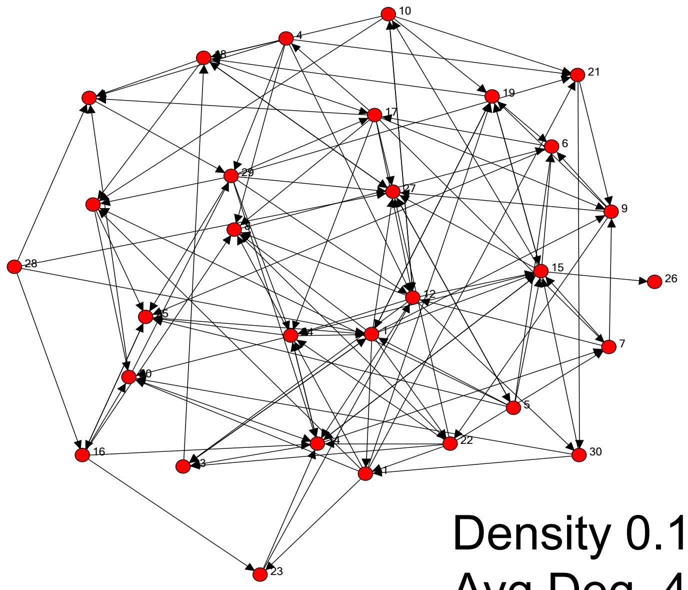
	1	2	3	4	5	6	7	8	9
1									
2	7%								
3	5%	0%							
4	5%	3%	2%						
5	0%	0%	6%	0%					
6	0%	0%	13%	0%	0%				
7	13%	2%	0%	3%	0%	11%			
8	0%	0%	7%	2%	6%	11%	14%		
9	9%	14%	0%	7%	0%	50%	0%	0%	

Average Degree

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



Density 0.47
Avg Deg 4



Density 0.14
Avg Deg 4

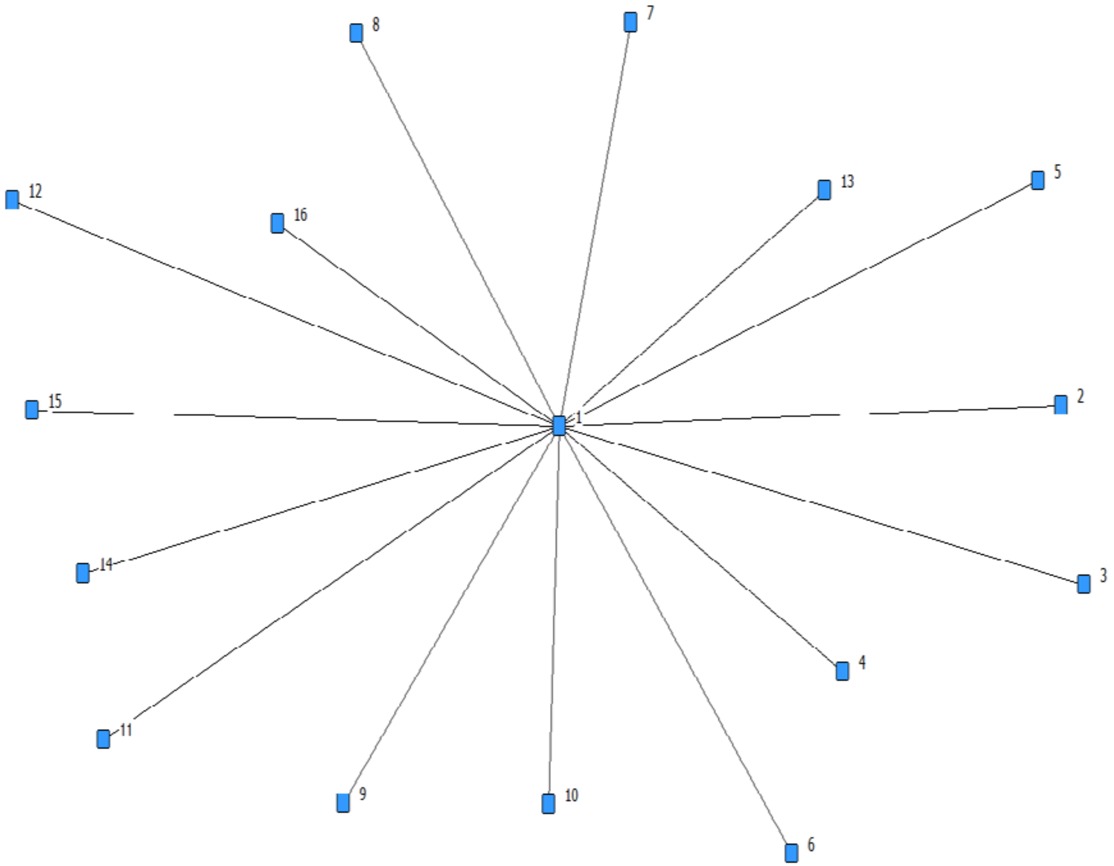
Degree variance and centralization

- **Variance** in degree (or any node level measure) indicates some people are much more central than others.
- Centralization is a kind of variance: the extent to which one person has all of the centrality
  - **Normal variance is variation around the mean**
    - i.e. sum of differences from the mean
  - **Centralization is variation around the maximum**
    - i.e. sum of differences (*squared*) from the maximum

id	Degree
17	14
16	11
7	5
15	5
13	4
1	3
8	3
9	3
11	3
14	3
2	2
3	2
4	2
5	2
6	2
10	2
12	2

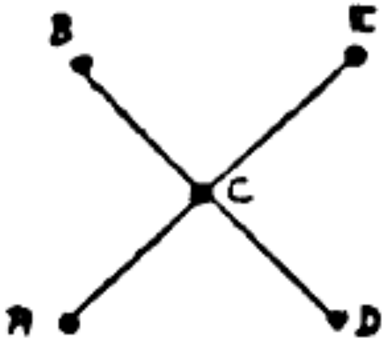
Centralization

- A network is maximally centralized with respect to any given node-level measure if the difference between the centrality of the most central node and that of all others is at a maximum
- For degree, it means the center is connected to all others, and they are only connected to the center

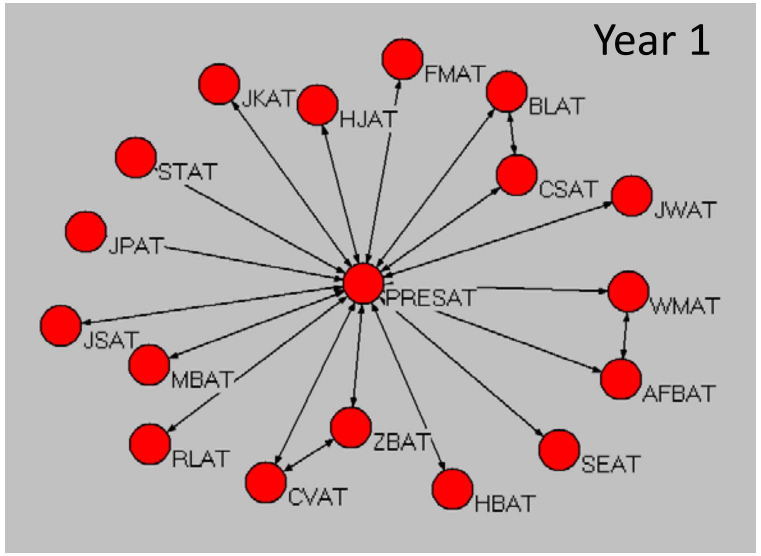


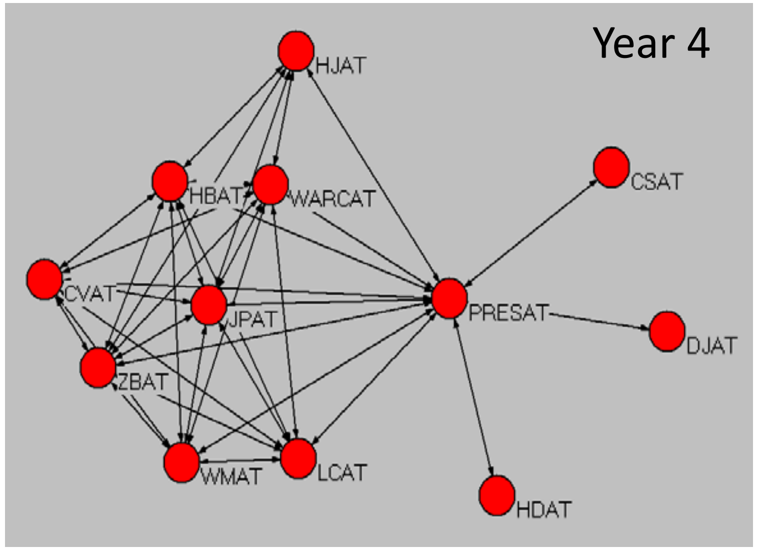
Calculating centralization

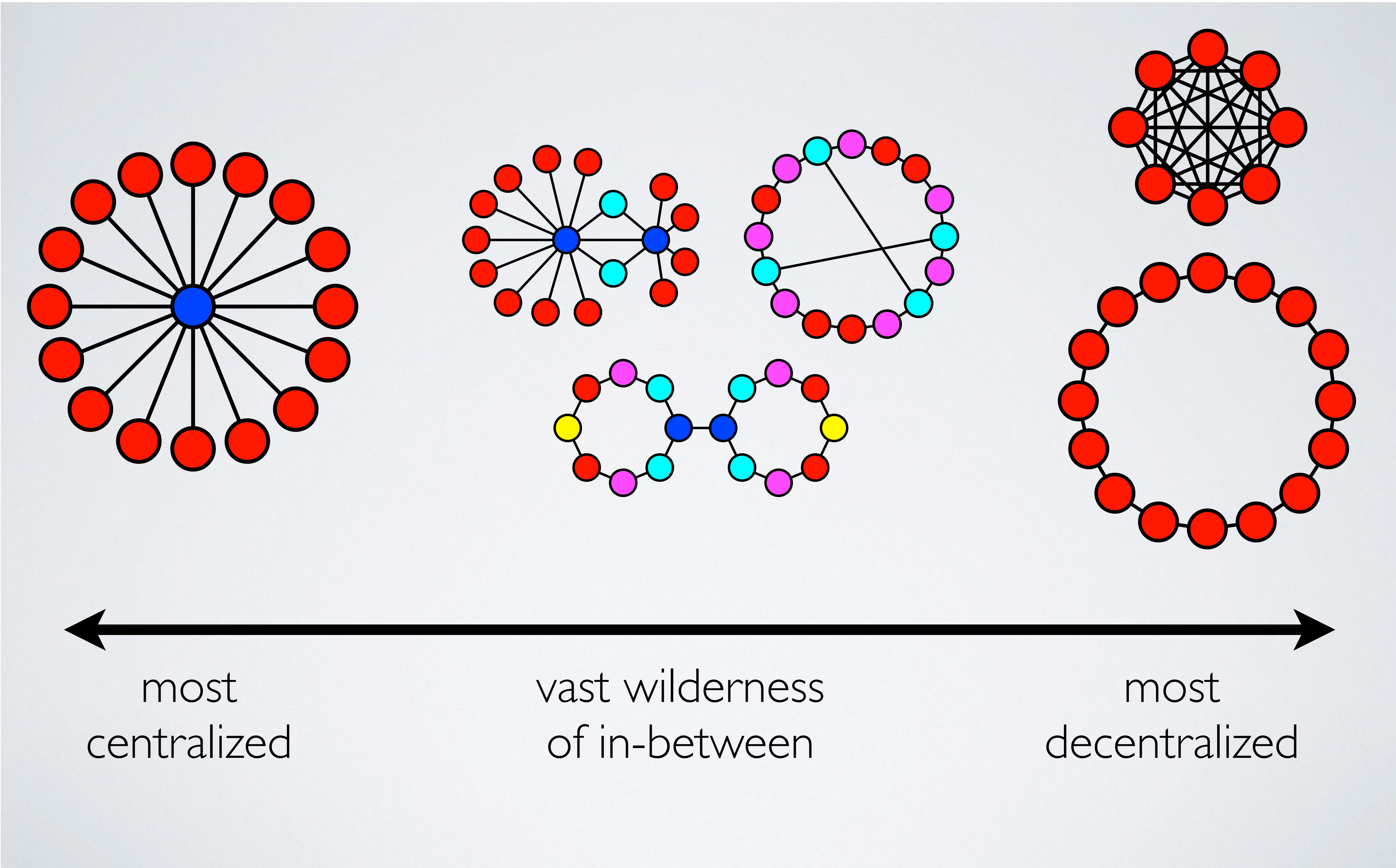
- Extent to which network revolves around a single node
- Sum of differences between the centrality of the most central node, and the centrality of every other node, divided by normalizing constant to make it run between 0 and 1
- Degree centralization:
    - $C = \frac{\sum_i d_{max} - d_i}{(n-1)(n-2)}$
    - (0+3+3+3+3)/(4*3) = 1.0



Carter admin.






most
centralized
vast wilderness
of in-between
most
decentralized

what have we learnt from it…

Baker & Faulkner (1993): Social Organization of conspiracy

(reconstructs communication networks in three well-known price-fixing conspiracies in the heavy electrical equipment industry to study social organization)

Questions: How are relations organized to facilitate illegal behavior?

Pattern of communication maximizes concealment, and predicts the criminal verdict.

Inter-organizational cooperation is common, but too much 'cooperation' can thwart market competition, leading to (illegal) market failure.

Illegal networks differ from legal networks, in that they must conceal their activity from outside agents. A "Secret society" should be organized to (a) remain concealed and (b) if discovered make it difficult to identify who is involved in the activity

The need for secrecy should lead conspirators to conceal their activities by creating ***sparse*** and ***decentralized*** networks.

- reconstructs communication networks in three well-known price-fixing conspiracies in the heavy electrical equipment industry to study social organization;
- findings:
    - structure of illegal networks is driven by need to maximize concealment, rather than efficiency;
    - structure also contingent on information-processing requirements;
    - person centrality in networks predicts *verdict, sentence* and *fine.*

Organization Objective	Information-Processing Requirement	
	High	Low
Concealment	$\boxed{\text{Centralized networks}}$	Decentralized networks
Coordination	Decentralized networks	Centralized networks

Figure 1. Concealment Versus Coordination: Theoretical Expectations and experimental results

Criminal Networks

- **Structure & Secrecy:**
  - Trade-off coordination and secrecy: maintain sufficient communication with minimal exposure;
  - Centralized structures facilitate communication, but increase exposure;
  - Decentralized/fragmented structures disperses information, increasing resilience, but hinder coordination;

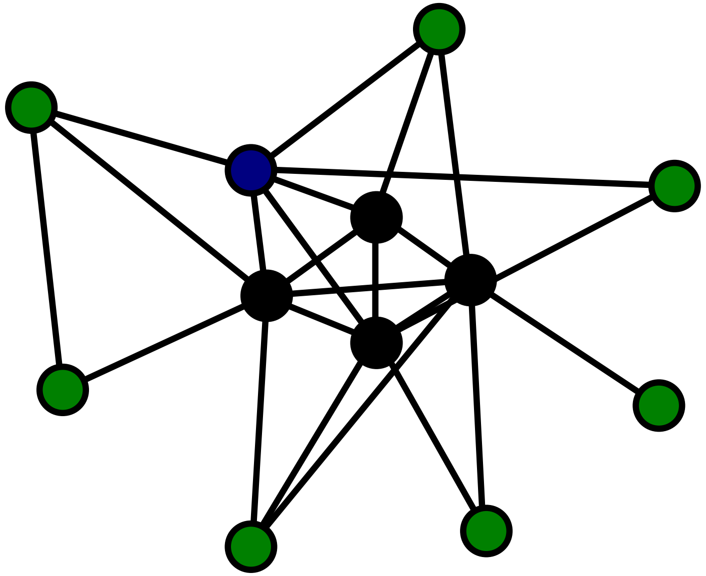
- **Position and role differentiation:**
  - intermediaries act as buffers and limit exposure of core members;

- **Redundancy and resilience:**
  - Redundant ties are used to build resilience and protect against disruption;
  - multiple people with overlaping roles;
  - reduces risk of single point of failure

- **Dynamic reconfiguration:**
  - shifts structure to avoid detection;
  - loosely coupled, flexible, able to adapt without collapsing

Core/Periphery

- Extent to which there is a “core” of people that holds the network together, such that
  - Core people are well connected to other core people, in general
  - Periphery people are connected to core people
  - Periphery people are NOT connected to other periphery people

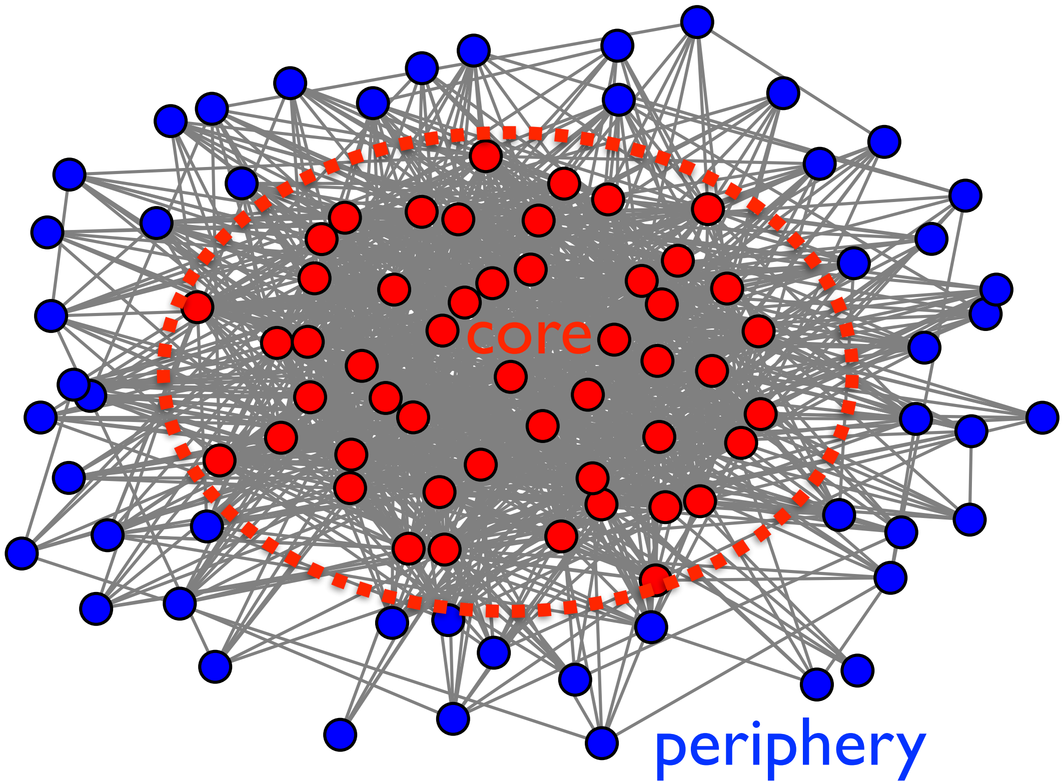


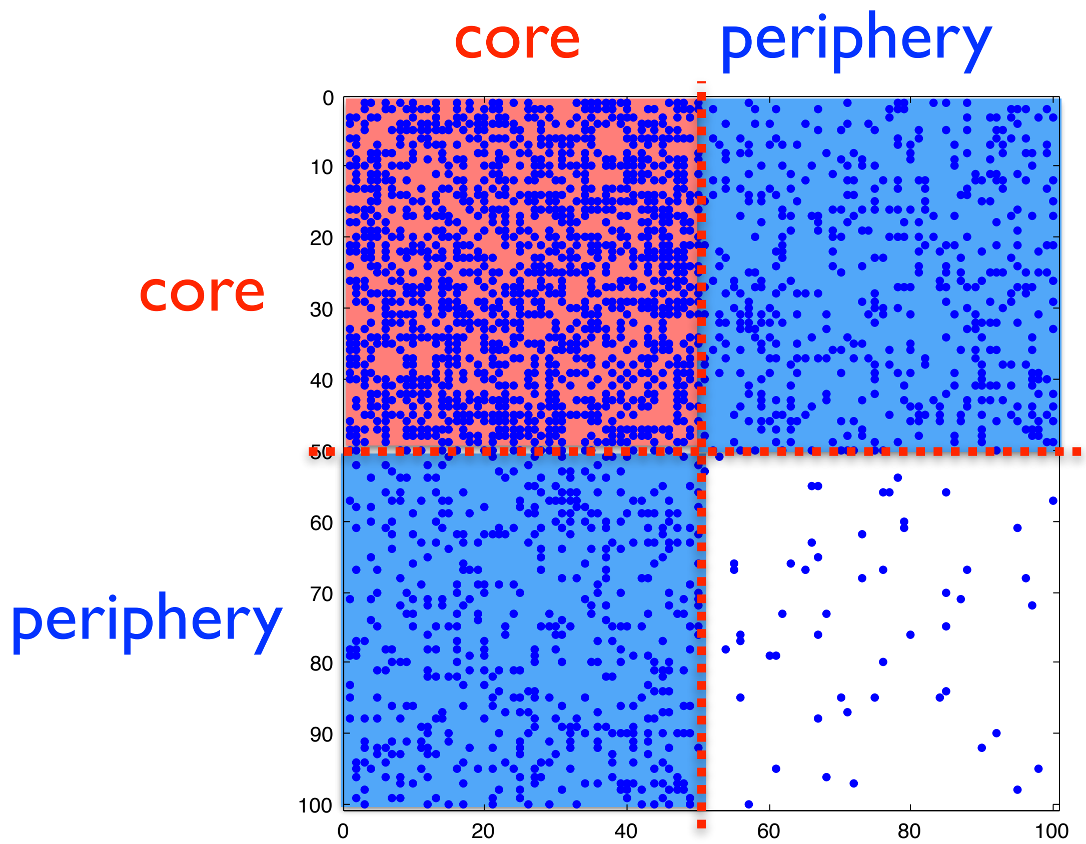
Core Periphery Block Model

Basic Idea:

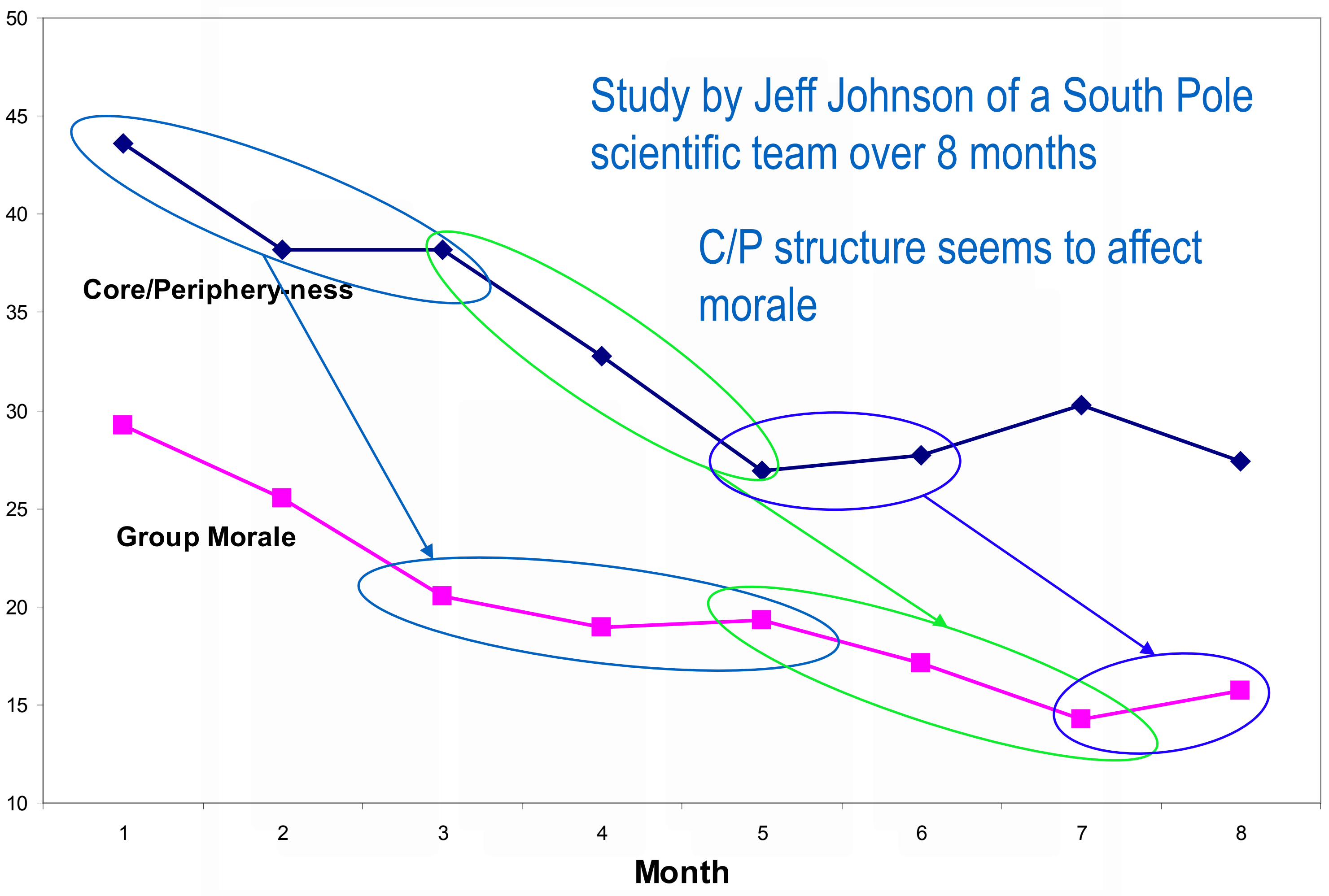
- A *module* or *community* is a collection of nodes defined by how its *edges* behave:
  - **Edge Density:** For social networks, we expect edge density to be greater within a community than without. (Assortative Community)
  - **Edge Weight:** For coexpression networks, we expect the correlations to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

Finding Core/Periphery Structures





C/P Structures & Morale



Kapferer tailor shop data

Bruce Kapferer (1972) observed interactions in a tailor shop in Zambia (then Northern Rhodesia) over a period of ten months. His focus was the changing patterns of alliance among workers during extended negotiations for higher wages.

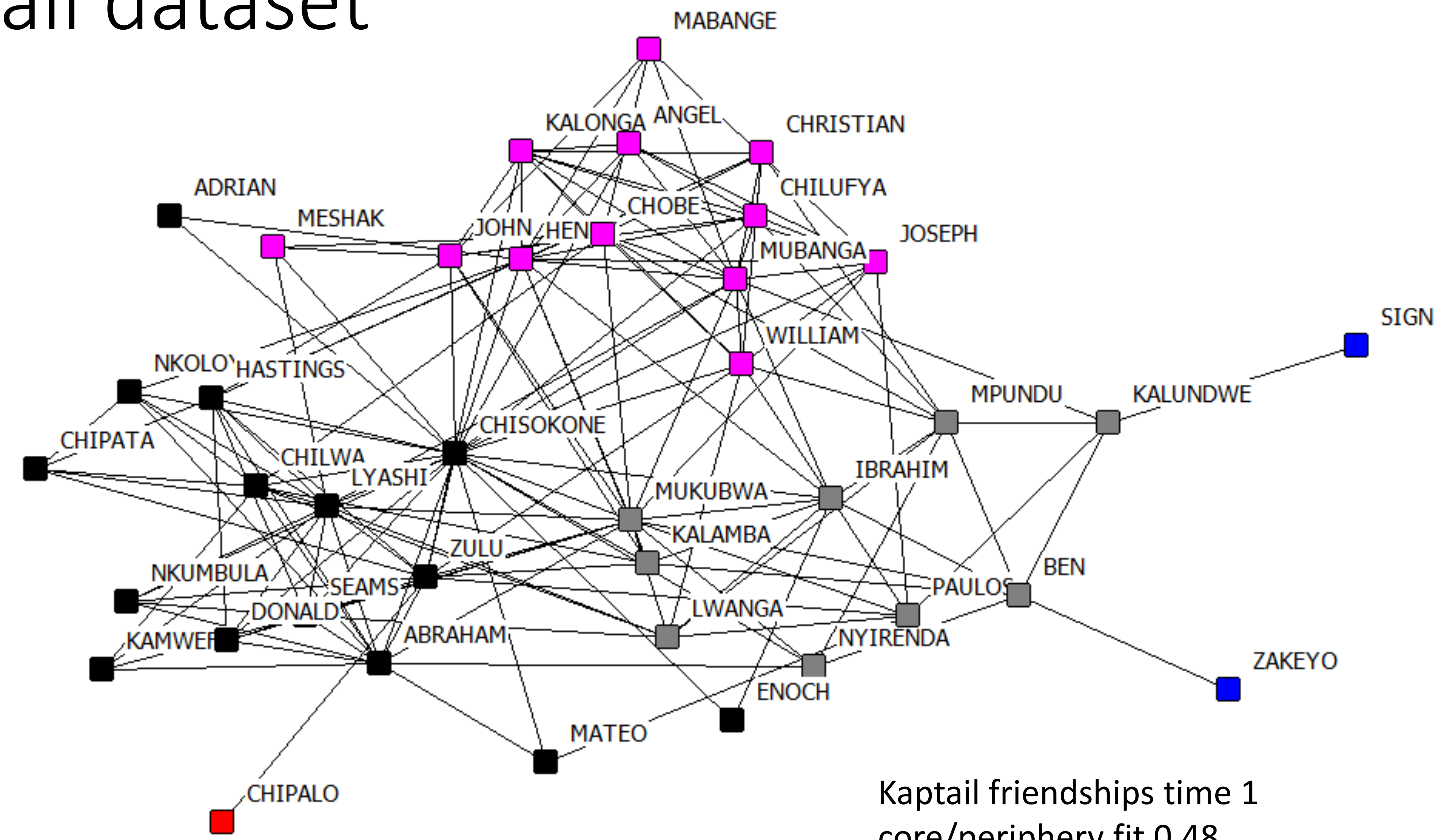
The matrices represent two different types of interaction, recorded at two different times (seven months apart) over a period of one month. TI1 and TI2 record the "instrumental" (work- and assistance-related) interactions at the two times; TS1 and TS2 the "sociational" (friendship, socioemotional) interactions.

The data are particularly interesting since an abortive strike occurred after the first set of observations, and a successful strike took place after the second.

-- UCINET help

Network|Core-periph|Categorical

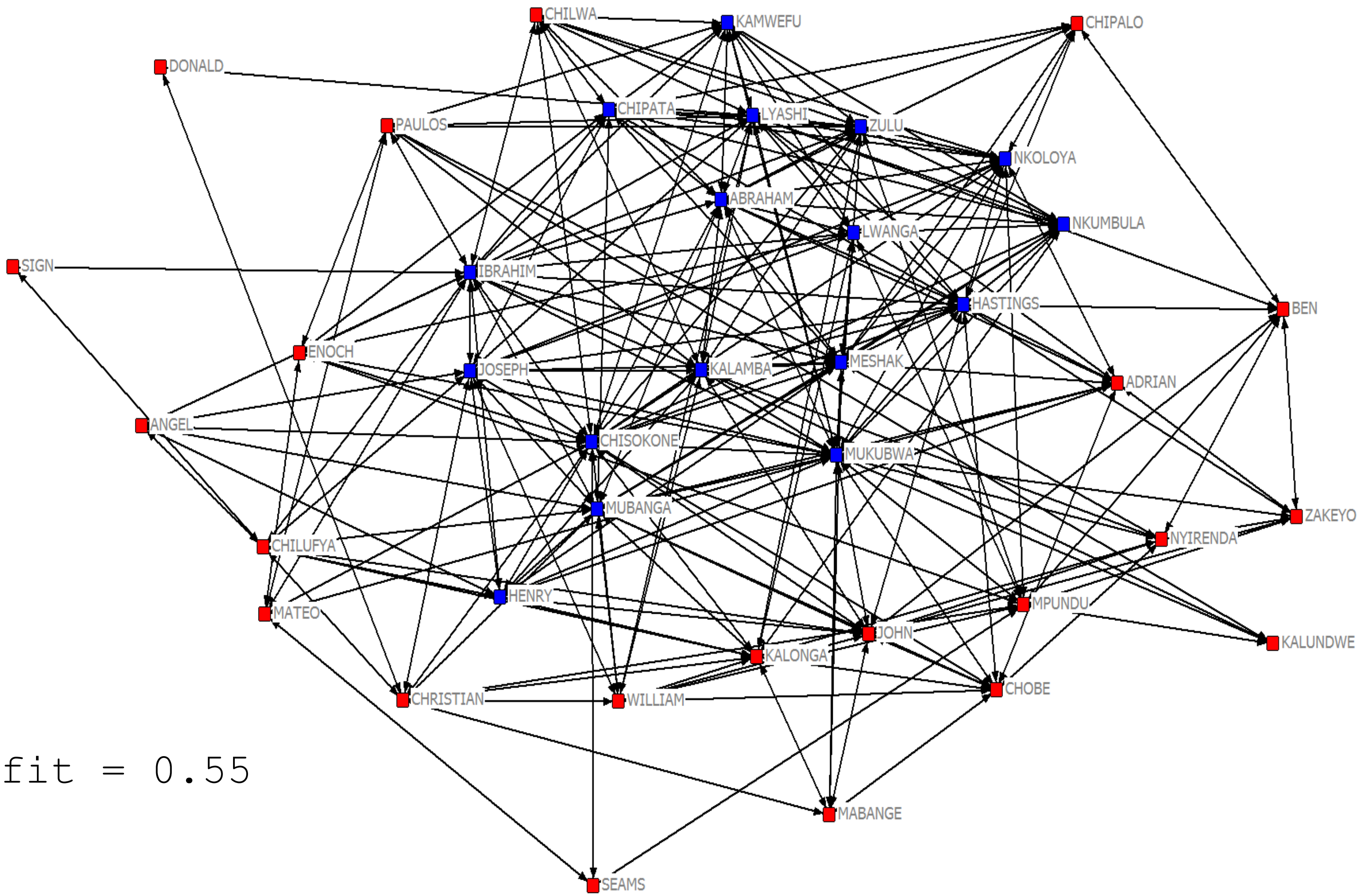
Kaptail dataset



Kaptail friendships time 1
core/periphery fit 0.48

Try cp on event by event matrix
Run Network|Cohesion|multiple measures ~kaptail

Kaptail time 2



Core/Periphery fit = 0.55

Finding Core/Periphery Structures

- Two approaches
  - Discrete/blockmodeling
    - Use combinatorial optimization to partition nodes into core and periphery sets such that core-core ties are maximized and periphery-periphery ties are minimized
  - Continuous
    - Calculate coreness of each node by modeling existence/strength of ties between pair of nodes as function of coreness of each

Categorical Approach

- Use combinatorial optimization to partition nodes into core and periphery sets such that
  - core-core ties are maximized
  - periphery-periphery ties are minimized
  - Core to Periphery: unspecified, but normally expect in-between value

$$\begin{pmatrix}
* & 1 & 1 & 1 & & & & & & \\
1 & * & 1 & 1 & & & & & & \\
1 & 1 & * & 1 & & & & & & \\
1 & 1 & 1 & * & 0 & 0 & 0 & 0 & 0 & 0 \\
 & & & 0 & * & 0 & 0 & 0 & 0 & 0 \\
 & & & 0 & 0 & * & 0 & 0 & 0 & 0 \\
 & & & 0 & 0 & 0 & * & 0 & 0 & 0 \\
 & & & 0 & 0 & 0 & 0 & * & 0 & 0 \\
 & & & 0 & 0 & 0 & 0 & 0 & * & 0 \\
 & & & 0 & 0 & 0 & 0 & 0 & 0 & *
\end{pmatrix}$$

Categorical Results

```
               1 2 3 3 5 5 7 2 1 9 1 6 4 4 2 4 0   9 4 8 7 8 0 0 6 2 7 8 9 6 1 3 3 5 5 6 7 8 9
               K N A H C M N Z K M L C M L H I J   C S M E P C S D Z K M J A W B C N C K A C M
               --------------------------------------------------------------------------------
 1   KAMWEFU |   1 1 1 1     1   1 1 1   1   1   | 1       1                                   |
 2  NKUMBULA | 1   1 1 1 1   1   1 1 1 1 1       |                                     1       |
 3   ABRAHAM | 1 1   1 1 1 1 1 1 1 1 1 1 1   1   | 1         1             1                   |
13  HASTINGS | 1 1 1   1 1 1   1 1 1 1       1 1 |           1     1     1 1   1 1             |
 5   CHIPATA | 1 1 1 1   1 1 1     1   1     1 1 | 1     1   1                                 |
25    MESHAK |   1 1 1 1   1   1 1   1 1 1 1 1 1 |         1 1             1       1         1 |
 7   NKOLOYA |     1 1 1 1   1   1 1 1   1     1 | 1         1         1   1                   |
12      ZULU | 1 1 1   1   1   1 1 1 1       1   | 1       1           1               1       |
21   KALAMBA |     1 1   1   1   1 1 1     1 1 1 | 1       1         1   1   1     1           |
19   MUKUBWA | 1 1 1 1   1 1 1 1   1 1 1 1 1 1 1 |     1 1         1 1 1 1 1     1 1         1 |
11    LYASHI | 1 1 1 1 1   1 1 1 1       1   1 1 | 1       1 1   1         1 1 1               |
16 CHISOKONE | 1 1 1 1   1 1 1 1 1         1 1 1 | 1 1 1 1               1   1   1   1 1 1     |
34   MUBANGA |   1 1   1 1       1         1 1 1 |                       1 1 1   1   1 1 1 1   |
14    LWANGA | 1 1 1     1 1     1 1         1 1 |       1             1               1       |
32     HENRY |           1     1 1   1 1     1 1 |                       1 1           1 1 1   |
24   IBRAHIM | 1   1 1 1 1   1 1 1 1 1 1 1 1   1 | 1   1 1 1   1                         1 1   |
30    JOSEPH |       1 1 1 1   1 1 1 1 1 1 1 1   |                           1       1   1 1   |
               --------------------------------------------------------------------------------
 9    CHILWA | 1   1   1   1 1 1   1 1       1   |                                             |
 4     SEAMS |                       1           |     1               1                       |
 8     MATEO |                   1   1       1   |   1   1 1                                   |
17     ENOCH |         1         1   1   1   1   |     1   1           1                       |
18    PAULOS | 1         1   1 1   1         1   |     1 1           1                         |
10   CHIPALO |     1 1 1 1 1       1             |                             1               |
20      SIGN |                               1   |                                         1   |
 6    DONALD |                     1             |                                   1         |
22    ZAKEYO |       1         1                 |                     1 1 1   1   1           |
27  KALUNDWE |                 1 1               |         1           1                       |
28    MPUNDU |             1 1   1       1       |   1   1         1 1       1 1     1         |
29      JOHN |       1         1 1   1 1   1     |                 1         1 1 1 1       1 1 |
26    ADRIAN |     1 1   1 1     1 1   1   1     |                 1             1             |
31   WILLIAM |                 1   1 1 1       1 |                     1 1       1   1 1       |
23       BEN |       1             1             |           1     1   1 1         1           |
33     CHOBE |       1         1   1 1           |                       1 1 1     1   1     1 |
15  NYIRENDA |           1     1 1               |                 1     1     1 1     1       |
35 CHRISTIAN |                       1 1       1 |               1     1     1         1   1 1 |
36   KALONGA |   1           1       1 1 1 1     |                           1   1 1 1     1 1 |
37     ANGEL |                       1 1   1 1 1 |                                         1   |
38  CHILUFYA |                         1   1 1 1 |             1         1           1 1 1     |
39   MABANGE |           1       1               |                       1       1   1 1       |
               --------------------------------------------------------------------------------
```

Density matrix

	1	2
1	0.699	0.235
2	0.235	0.173

Kaptail-kapfts2

Continuous approach

- Discrete model effectively creates binary coreness variable such that ties between i and j are given by product of coreness of each
  - If ci and cj = 1 then Xij = 1
  - If ci = 1 and cj = 0, then Xij = 0
  - if ci and cj = 0 then Xij = 0
- So this could be generalized to real-valued coreness vector

core ness		1	1	1	0	0	0	0
		a	b	c	d	e	f	g
1	a	1	1	1	0	0	0	0
1	b	1	1	1	0	0	0	0
1	c	1	1	1	0	0	0	0
0	d	0	0	0	0	0	0	0
0	e	0	0	0	0	0	0	0
0	f	0	0	0	0	0	0	0
0	g	0	0	0	0	0	0	0

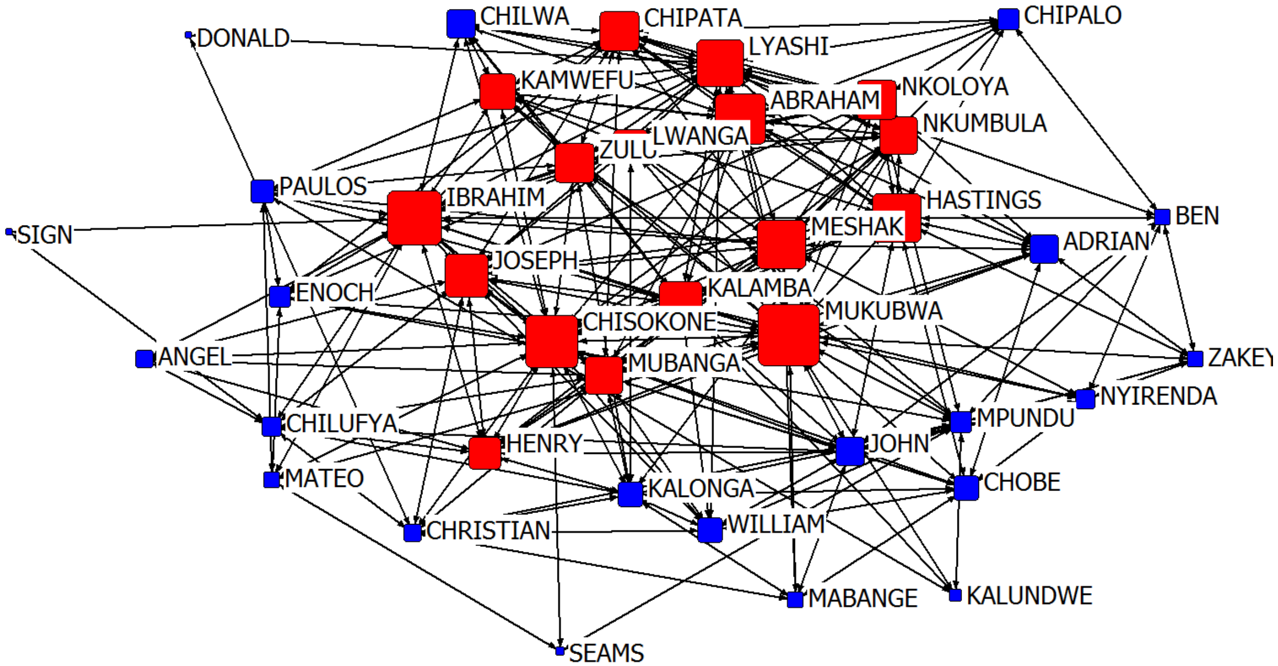
Continuous approach

- We generalize to continuous coreness scores such that prob/strength of a tie between I and j is a function of the coreness of each
  - Xij = f(ci*cj)
  - If both have high coreness, then tied to each other
  - If both have low coreness, then not tied
- We use a least-squares type procedure to find scores c to minimize

$$\sum_{i,j}(x_{ij} - c_i c_j)^2$$

- Fitting a model of ties
  - Could use r-square to measure fit of model

Continuous coreness



Colors based on the discrete model. Sizes based on continuous model

		1 Corene
16	CHISOKONE	0.406
19	MUKUBWA	0.304
11	LYASHI	0.249
34	MUBANGA	0.242
32	HENRY	0.233
12	ZULU	0.232
3	ABRAHAM	0.213
13	HASTINGS	0.184
30	JOSEPH	0.182
24	IBRAHIM	0.181
31	WILLIAM	0.174
4	SEAMS	0.173
36	KALONGA	0.160
21	KALAMBA	0.157
38	CHILUFYA	0.157
29	JOHN	0.152
6	DONALD	0.143
33	CHOBE	0.142
9	CHILWA	0.141
14	LWANGA	0.128
35	CHRISTIAN	0.125
37	ANGEL	0.124
7	NKOLOYA	0.119
2	NKUMBULA	0.114
18	PAULOS	0.102
28	MPUNDU	0.101
15	NYIRENDA	0.099
39	MABANGE	0.085
25	MESHAK	0.085
5	CHIPATA	0.082
23	BEN	0.080
1	KAMWEFU	0.069

Measure cpness

- Both discrete and continuous approaches fit a model to the data, i.,e., predict ties
  - Discrete
    - If ci = 1 and cj = 1 then xij = 1
    - If ci = 0 and cj = 0 then xij = 0
  - Continuous
    - Prob(xij) = f(ci*cj)
- So in both cases we can measure goodness of fit
  - Degree to which data conforms to idealized cp structure

Reciprocity

MAN convention:
- Mutuals
- Asymmetrics
- Nulls

- Let R = number of reciprocated arcs, U = number of unreciprocated arcs
- Arc reciprocity
  - Proportion of outgoing ties that are answered with an incoming tie
  - R/(R+U)
- Dyad reciprocity
  - Proportion of non-null dyads that are symmetric ("mutuals")
  - R/(R+2U)

```
Reciprocity measures CAMPNET
                           ------

   1          Recip Arcs      38
   2        Unrecip Arcs      16
   3            All Arcs      54
   4     Arc Reciprocity   0.704
   5           Sym Dyads      19
   6          Asym Dyads      16
   7     All ~null Dyads      35
   8    Dyad Reciprocity   0.543
```

Calculating Reciprocity

- Dyad Method

$$\frac{\#Reciprocated\ Dyads}{\#Adjacent\ Dyads}$$

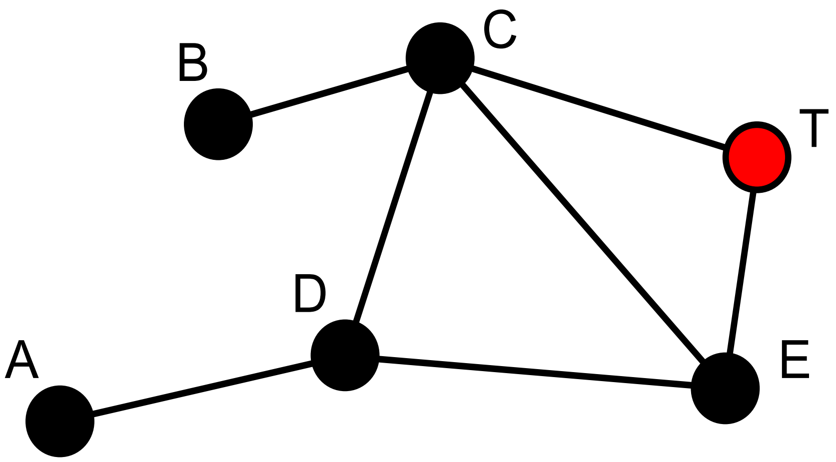
- Arc Method

$$\frac{\#Reciprocated\ Arcs}{\#Total\ Arcs}$$

- Hybrid methods
  - When partitioned: uses Arc Method between groups and Dyad Method within groups
    - When not partitioned, same as Dyad Method

Transitivity

- Proportion of triples with 3 ties as a proportion of triples with 2 or more ties
  - Aka the wtd clustering coefficient
- A clumpiness measure?



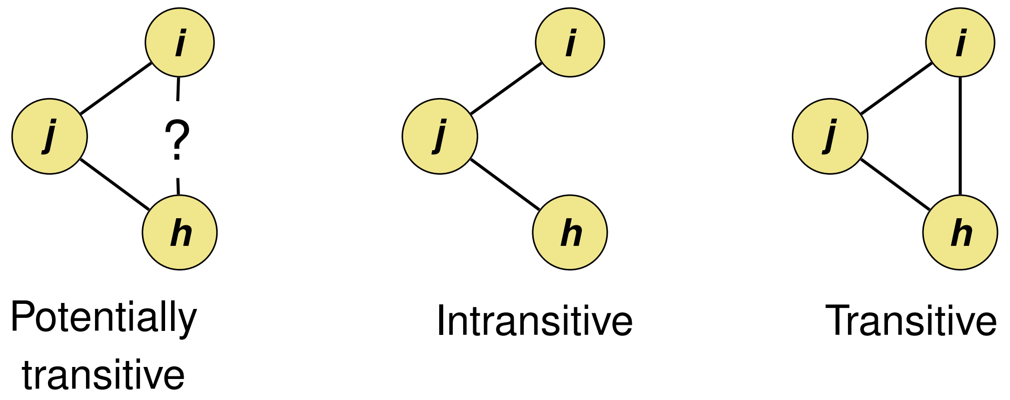
cc = 12/26 = 46.15%

{C,T,E} is a transitive triple, but {B,C,D} is not. {A,D,T} is not counted at all.

Transitivity

The tendency for a tie from i to k to occur at greater than chance frequencies if there are ties from i to j and from j to k – the i to j tie completes “transitively” the triple consisting of the tie from i to j and the tie from j to k.

Transitivity depends on *triads*, subgraphs formed by 3 nodes



measuring transitivity – **clustering index**

A measure for transitivity is the (global) transitivity index, defined as the ratio

$$\text{Transitivity Index} \;=\; \frac{\sharp\text{Transitive triads}}{\sharp\text{ Potentially transitive triads}}\;.$$

(Note that "$\sharp A$" means the number of elements in the set A.) This also is sometimes called a *clustering* index.

This is between 0 and 1; it is 1 for a transitive graph.
For random graphs, the expected value of the transitivity index is close to the density of the graph (why?);
for actual social networks,
values between 0.3 and 0.6 are quite usual.

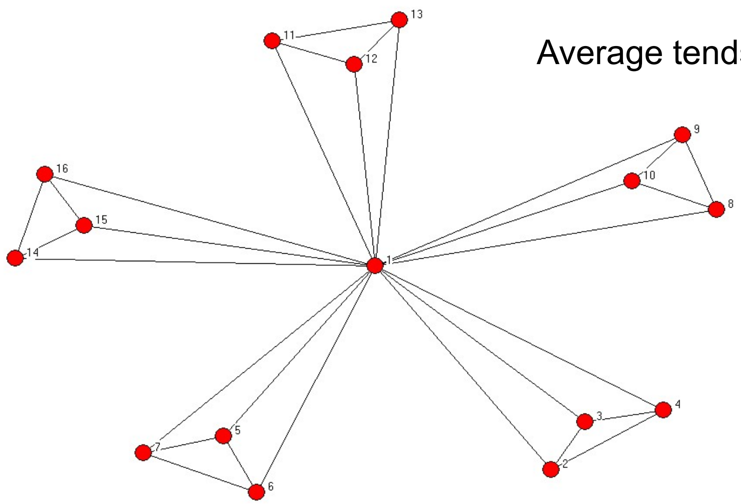
Clustering

What fraction of my friends are friends of each other?

(1) Calculate clustering for a particular node;

(1) Average individual clustering coefficients across the network (it weights clustering node by node)

(2) Overall clustering: out of all possible triplets in the network, what the frequency with which it i



Average tends to 1

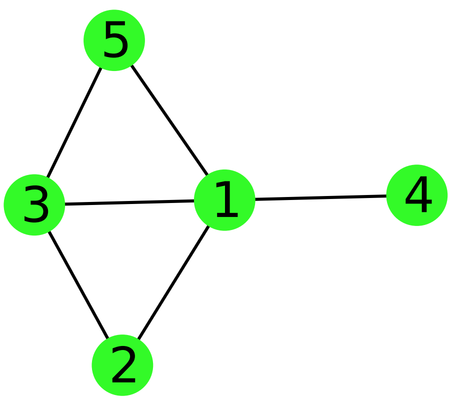
Overall tends to 0

local clustering coefficient

If i is a node with $k_i \geq 2$ then its *local clustering coefficient* is defined as:

$$C_i = \frac{\text{Number of triangles containing } i}{\text{Number of pairs of neighbours of } i},$$
$$= \frac{t_i}{\frac{1}{2}k_i(k_i - 1)},$$

where $t_i = [A^3]_{ii}$.



Possible triangles including node 1:

$$\{(1-2-3), (1-3-5), (1-2-5),$$
$$(1-5-4), (1-2-4), (1-3-4)\}.$$

Actual triangles:

$$\{(1-2-3), (1-3-5)\}.$$

$$C_1 = \frac{1}{3}.$$

global clustering coefficient

There are two alternative definitions of the global clustering coefficient:

Version 1: Average Clustering Coefficient

$$C = \langle C_i \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i.$$

Version 2: Overall Clustering Coefficient

$$C = \frac{3 \times t}{\text{number of connected triples}}$$

where t is the total number of triangles. If there are no self-loops then $t = \frac{1}{3}\text{trace}(A^3)$.

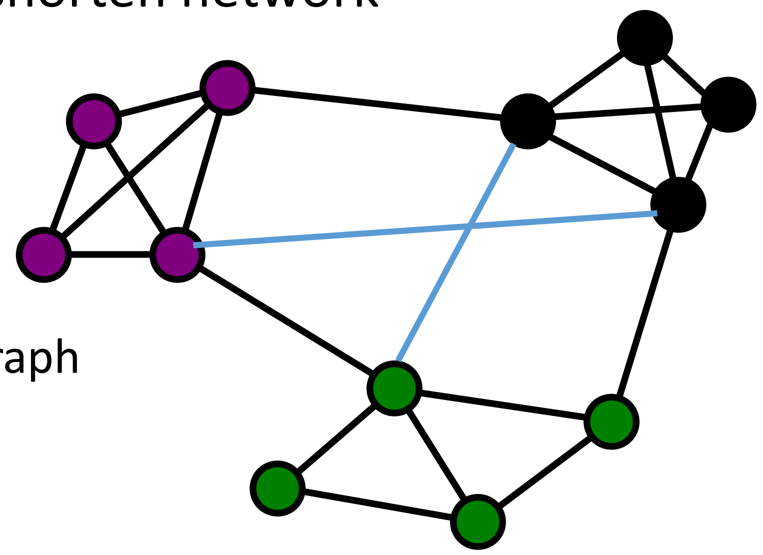
Notes on Clustering Coef

- Unweighted measure
  - Node level clustering coefficient (cc_i) For each node, measure density of their ego network (not including ego)
  - Average cc_i for all *i* to get overall network-level clustering coef
  - Seen as a measure of clumpiness
- Weighted measure
  - When averaging, weight each node by the number of pairs of alters in neighborhood
  - This value is precisely equal to transitivity

Small Worldness

$$\sigma = \frac{C(A)/C(R)}{L(A)/L(R)}$$

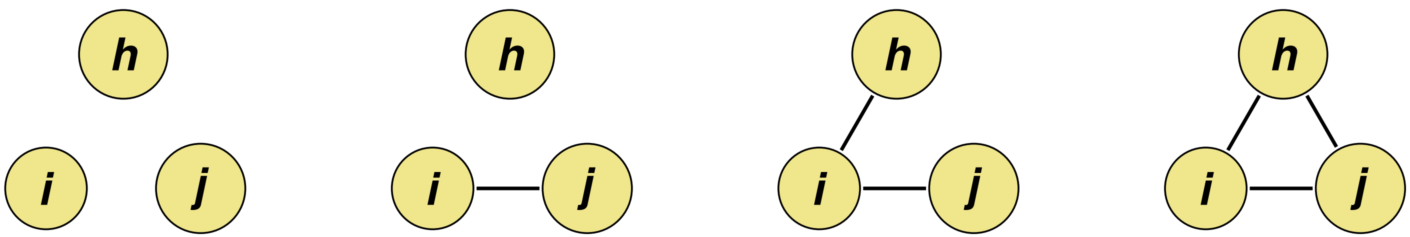
- Theory
  - Human networks typically clumpy
    - Homophily, balance theory, temporal-spatial opportunities
  - In the space of all possible graphs, clumpy graphs tend to have longer distances
    - But as milgram seemed to show, human networks have short distances
  - Watts and Strogatz: a very few random ties will radically shorten network
- Method
  - A network is a small world if it is both clumpy and has short distances
  - How clumpy is clumpy? How short is short? Comparison with random graphs
    - C(A) = clust coef of actual graph; C(R) = clus coef of random graph
    - L(A) = avg dist in actual graph; L(R) = avg dist in random graph
  - Small worldness indices such as σ



local structure and triad counts

The studies about transitivity in social networks led Holland and Leinhardt (1975) to propose that the *local structure* in social networks can be expressed by the *triad census* or *triad count*, the numbers of triads of any kinds.

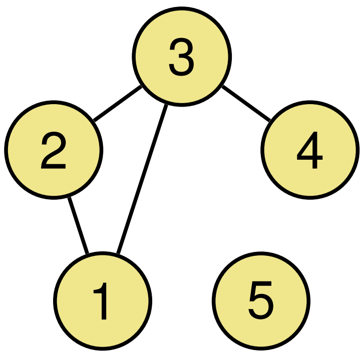
For (nondirected) graphs, there are four triad types:



Empty | One edge | Two-path / Two-star | Triangle

local structure and triad counts

A simple example graph
with 5 nodes.

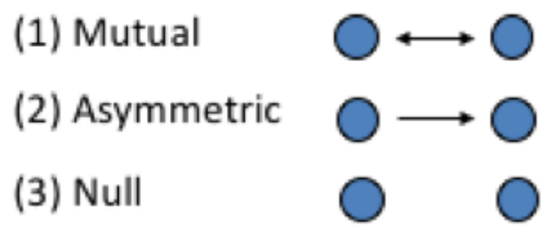


i	j	h	triad type
1	2	3	triangle
1	2	4	one edge
1	2	5	one edge
1	3	4	two-star
1	3	5	one edge
1	4	5	empty
2	3	4	two-star
2	3	5	one edge
3	4	5	one edge

In this graph, the triad census is $(1, 5, 2, 1)$
(ordered as: empty – one edge – two-star – triangle).

MAN coding for triad census

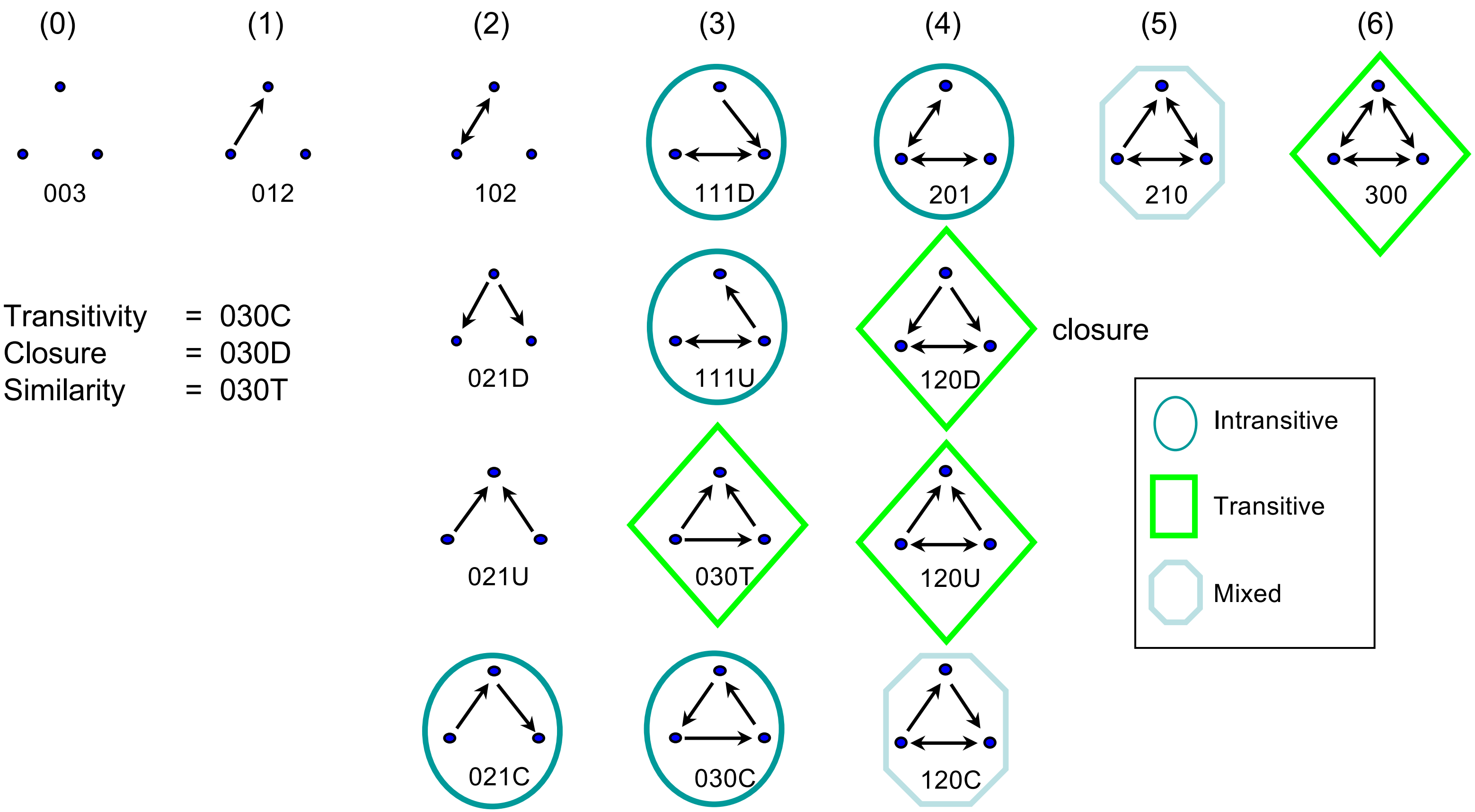
Holland and Leinhardt (1975) proposed the following MAS coding.



the scheme a further identifying letter: Up, Down, Cyclical, Transitive.

E.g. 120 has 1 mutual, 2 asymmetric, 0 null dyads and the Down orientation

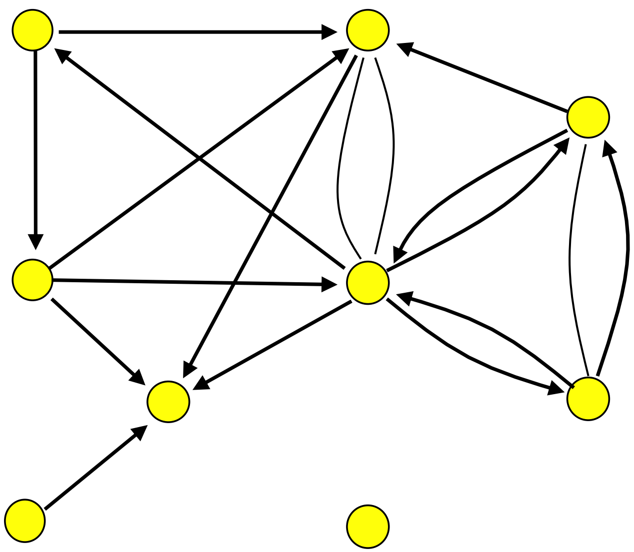
triad census



Transitivity = 030C
Closure = 030D
Similarity = 030T

Transitivity: tie **i** to **k** to occur if ties from **i** to **j** and **j** to **k** exist;
Closure: tie **i** to **j** to occur if persons **k** with ties to both **i** and **j** exist;
Similarity: tie **i** to **j** to occur if persons **k** to whom **i** and **j** have ties exist;

triad census - example



Type	Number of triads
1 - 003	21
2 - 012	26
3 - 102	11
4 - 021D	1
5 - 021U	5
6 - 021C	3
7 - 111D	2
8 - 111U	5
9 - 030T	3
10 - 030C	1
11 - 201	1
12 - 120D	1
13 - 120U	1
14 - 120C	1
15 - 210	1
16 - 300	1
Sum (2 - 16):	63

- triads define behavioral mechanisms: we can leverage the distribution of triads in a network to test whether the hypothesized mechanism is active.

- How?

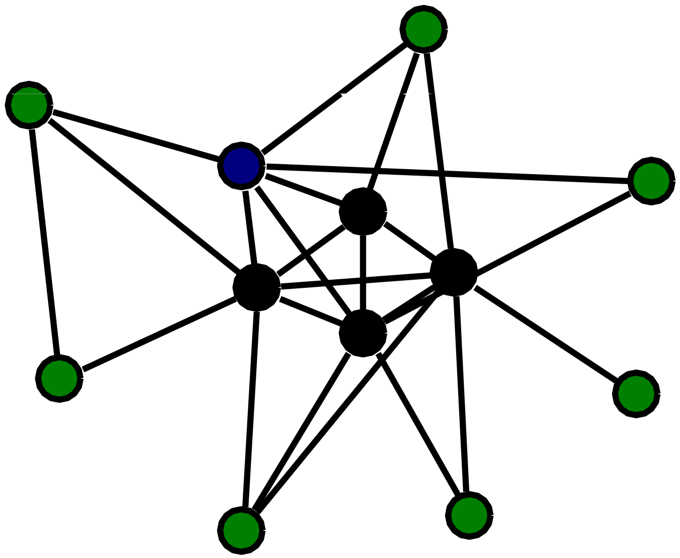
(1) Count the number of each triad type in a given network

(2) Compare to the expected number, given some (random) distribution of ties in the network;

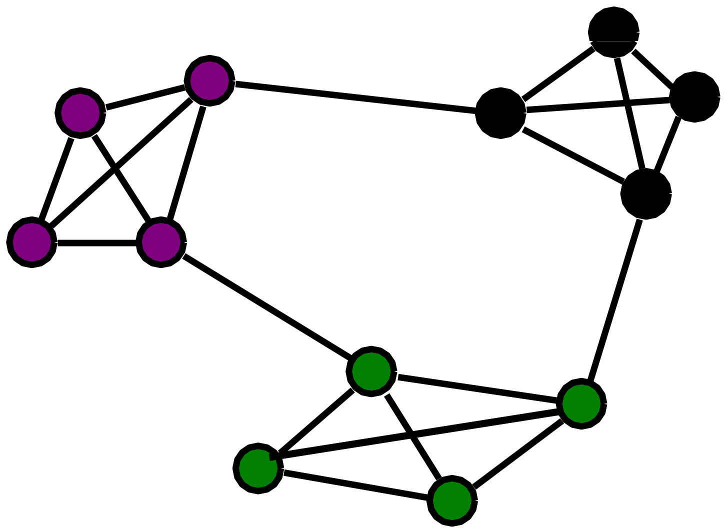
- Statistical approach proposed by Holland and Leinhardt is now obsolete. Statistical methods have been proposed for probability distributions of graphs depending primarily on triad counts, but complemented with stat counts and nodal variables, along with some higher-order configurations essential for adequate modeling of empirical network data.

Average Distance

- Average geodesic distance between all pairs of nodes



avg. dist. = 1.9



avg. dist. = 2.4

Average geodesic distance

Only makes sense in connected graphs

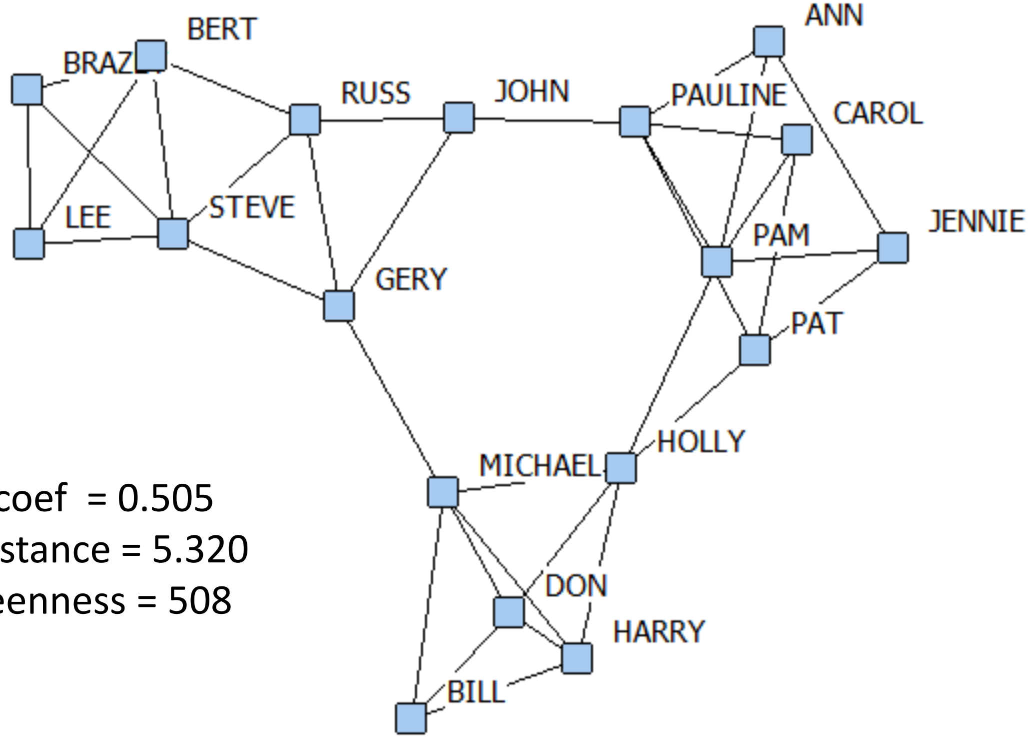
	HOLLY	BRAZEY	CAROL	PAM	PAT	JENNIE	PAULINE	ANN	MICHAEL	BILL	LEE	DON	JOHN	HARRY	GERY	STEVE	BERT	RUSS
HOLLY	0	4	2	1	1	2	2	2	1	2	4	1	3	1	2	3	4	3
BRAZEY	4	0	5	5	5	6	4	5	3	4	1	4	3	4	2	1	1	2
CAROL	2	5	0	1	1	2	1	2	3	4	5	3	2	3	3	4	4	3
PAM	1	5	1	0	2	1	1	1	2	3	5	2	2	2	3	4	4	3
PAT	1	5	1	2	0	1	1	2	2	3	5	2	2	2	3	4	4	3
JENNIE	2	6	2	1	1	0	2	1	3	4	6	3	3	3	4	5	5	4
PAULINE	2	4	1	1	1	2	0	1	3	4	4	3	1	3	2	3	3	2
ANN	2	5	2	1	2	1	1	0	3	4	5	3	2	3	3	4	4	3
MICHAEL	1	3	3	2	2	3	3	3	0	1	3	1	2	1	1	2	3	2
BILL	2	4	4	3	3	4	4	4	1	0	4	1	3	1	2	3	4	3
LEE	4	1	5	5	5	6	4	5	3	4	0	4	3	4	2	1	1	2
DON	1	4	3	2	2	3	3	3	1	1	4	0	3	1	2	3	4	3
JOHN	3	3	2	2	2	3	1	2	2	3	3	3	0	3	1	2	2	1
HARRY	1	4	3	2	2	3	3	3	1	1	4	1	3	0	2	3	4	3
GERY	2	2	3	3	3	4	2	3	1	2	2	2	1	2	0	1	2	1
STEVE	3	1	4	4	4	5	3	4	2	3	1	3	2	3	1	0	1	1
BERT	4	1	4	4	4	5	3	4	3	4	1	4	2	4	2	1	0	1
RUSS	3	2	3	3	3	4	2	3	2	3	2	3	1	3	1	1	1	0

	Geo Dist
Mean	2.66
Std Dev	1.26
Sum	814
Variance	1.60
SSQ	2654
MCSSQ	488.65
Euc Norm	51.52
Minimum	1
Maximum	6
N of Obs	306

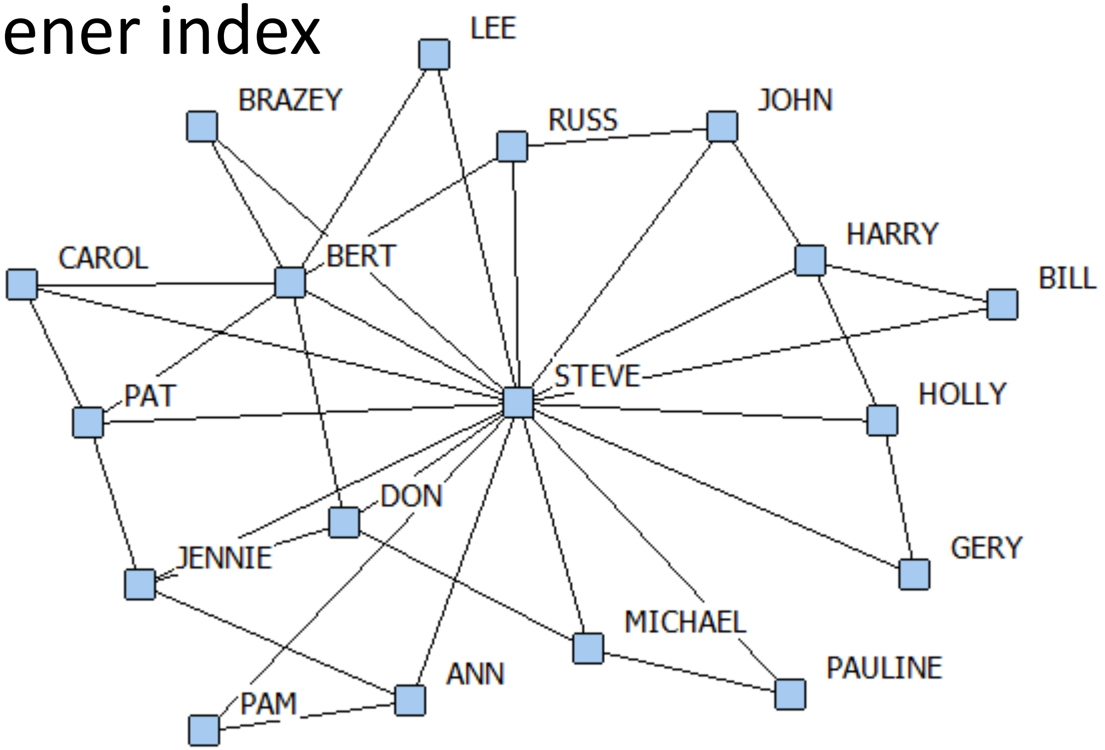
Average Distance

Clumpy networks tend to have longer distances

- Average geodesic distance between all pairs of nodes
- Sum of distances is known as the Wiener index



Clustering coef = 0.505
Avg Geo Distance = 5.320
Sum Betweenness = 508

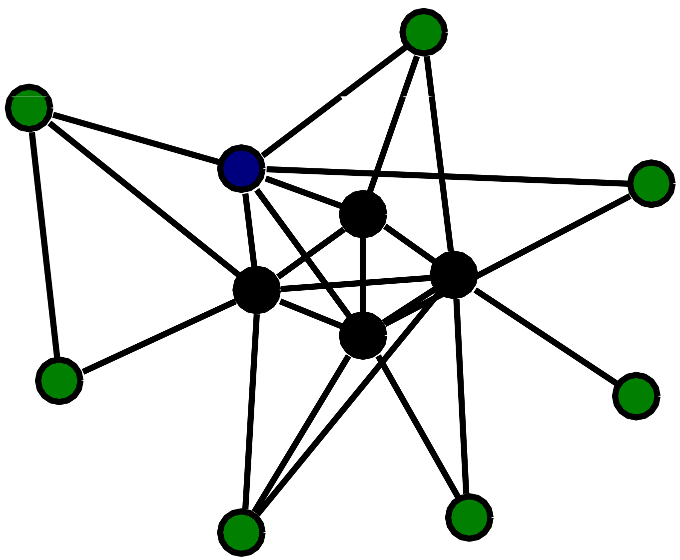


Clustering coef = 0.278
Avg Geo Distance = 3.542
Sum of betweenness = 236

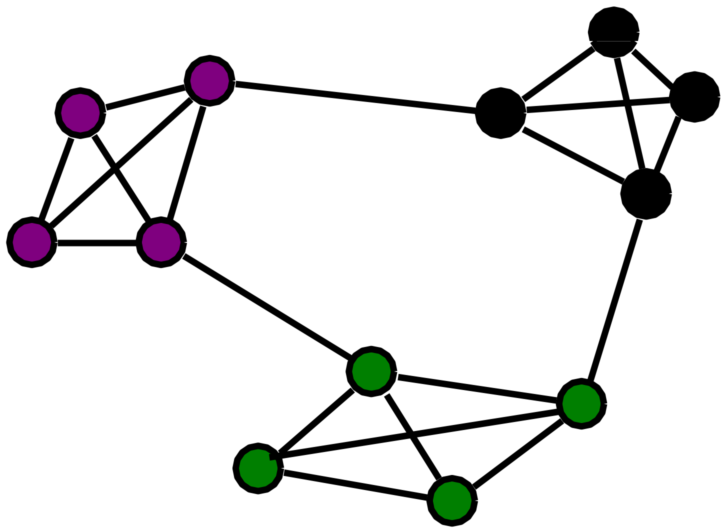
Note that the number of nodes and ties is th works

Diameter

- Maximum distance



Diameter = 3



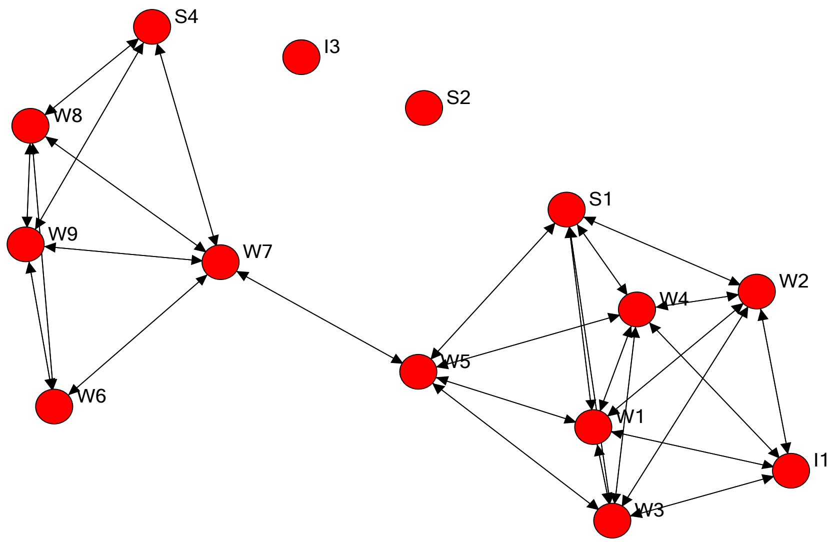
Diameter = 3

Fragmentation Measures

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

Component Ratio (CR)

- No. of components minus 1 divided by number of nodes minus 1



Measure of disconnectedness

CR = (3-1)/(14-1) = 0.154

CR is 1 when all nodes are isolates
CR is 0 when all nodes in one component

F Measure of Fragmentation

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

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

proportion of pairs of nodes that can reach each other via path.

Subtract from 1 to get proportion of pairs that **cannot** reach each other

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

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

Shortcut Formula for F Measure

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

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

S_k = size of kth component

Distance-Weighted Fragmentation

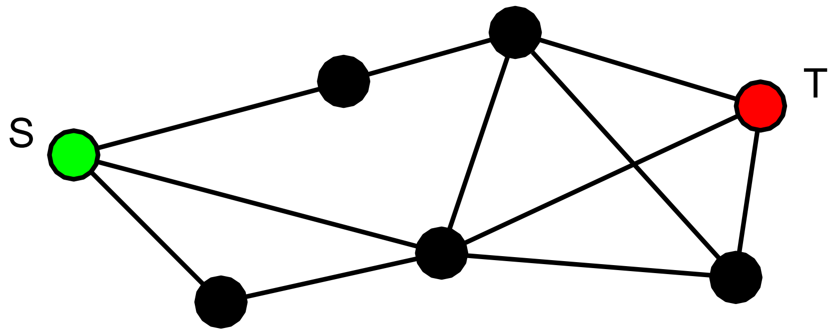
- Use the reciprocal of distance
  - letting $1/\infty = 0$

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

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

Connectivity

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



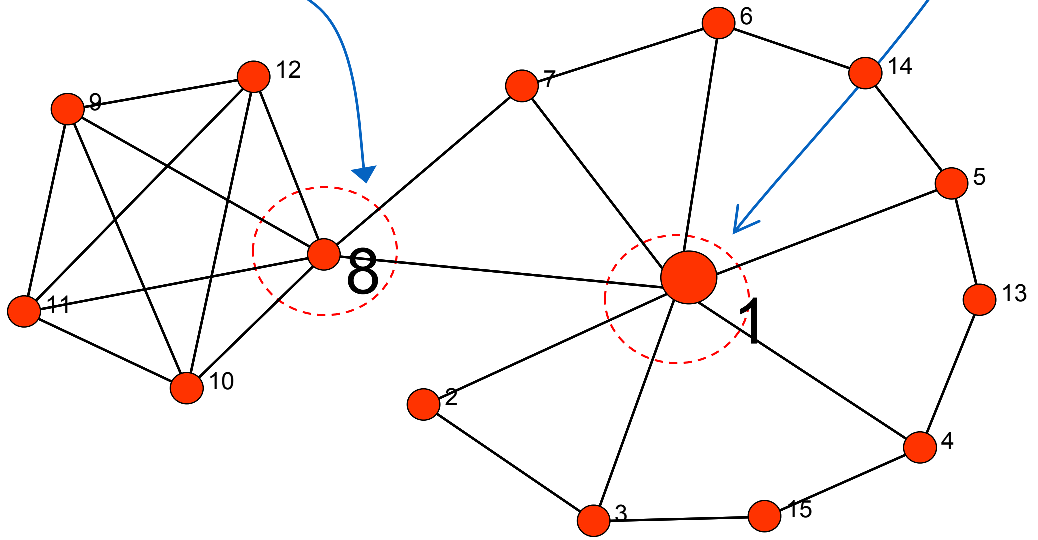
KeyPlayer application

- Suppose you want to disrupt a network
  - E.g., stop epidemic by immunizing/quarantining an affordable # of people
  - Disrupt terrorist group's ability to coordinate
- You have the resources to neutralize just k nodes. Which ones do you pick?
- Obvious solution is the pick the k most central nodes
- Two problems
  - Off-the-shelf measures are not designed for this specific purpose (but we can improvise) *Design Problem*
  - Picking an optimal set of k nodes is not the same thing as picking the k nodes that individually most optimal *Ensemble Problem*

DISRUPTION

The Design Issue

- By standard off-the-shelf measures of node centrality, node 1 is the most important player, but deleting it …
  - does not disconnect the network
- In contrast, deleting node 8 breaks network into two components
  - Yet node 8 is not highest in centrality
- Standard off-the-shelf centrality measures not optimal for the purpose of disrupting networks
  - Nor many other specific purposes

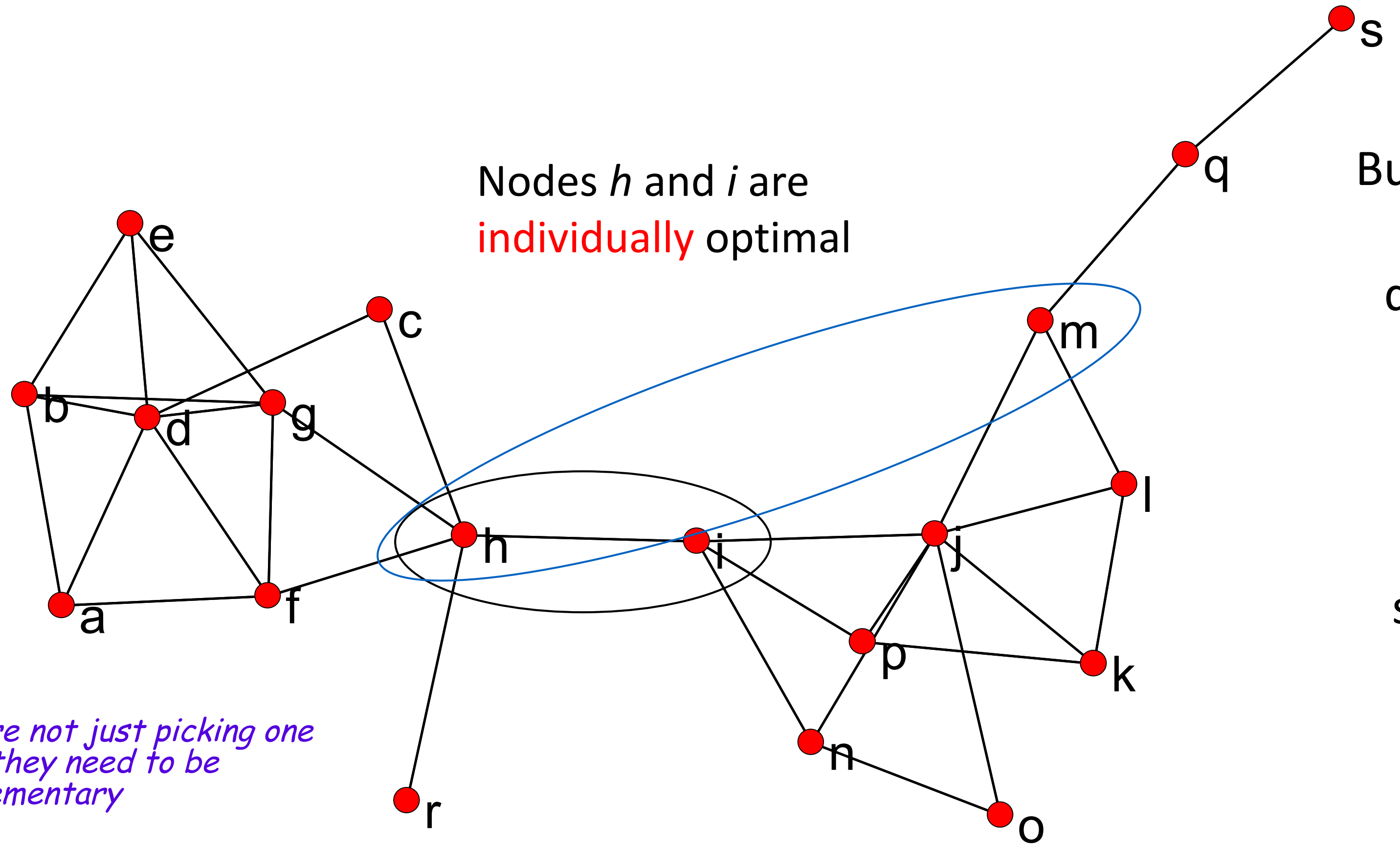


DISRUPTION

The Ensemble Issue

Structural redundancy creates need for choosing complementary nodes

Nodes *h* and *i* are individually optimal



But deleting both is no better than deleting *h* alone -- *h* and *i* are redundant

In contrast, {*h,m*} splits graph into 4 fragments (is optimal)

if we're not just picking one node, they need to be complementary

- Choosing optimal **set** of *k* players is not same as choosing the *k* best players

KeyPlayer – cont.

- Use a combinatorial optimization algorithm to identify the best combination of k nodes to remove
- Measure "bestness" of a particular combination by the amount of increase in fragmentation as measured by F or breadth

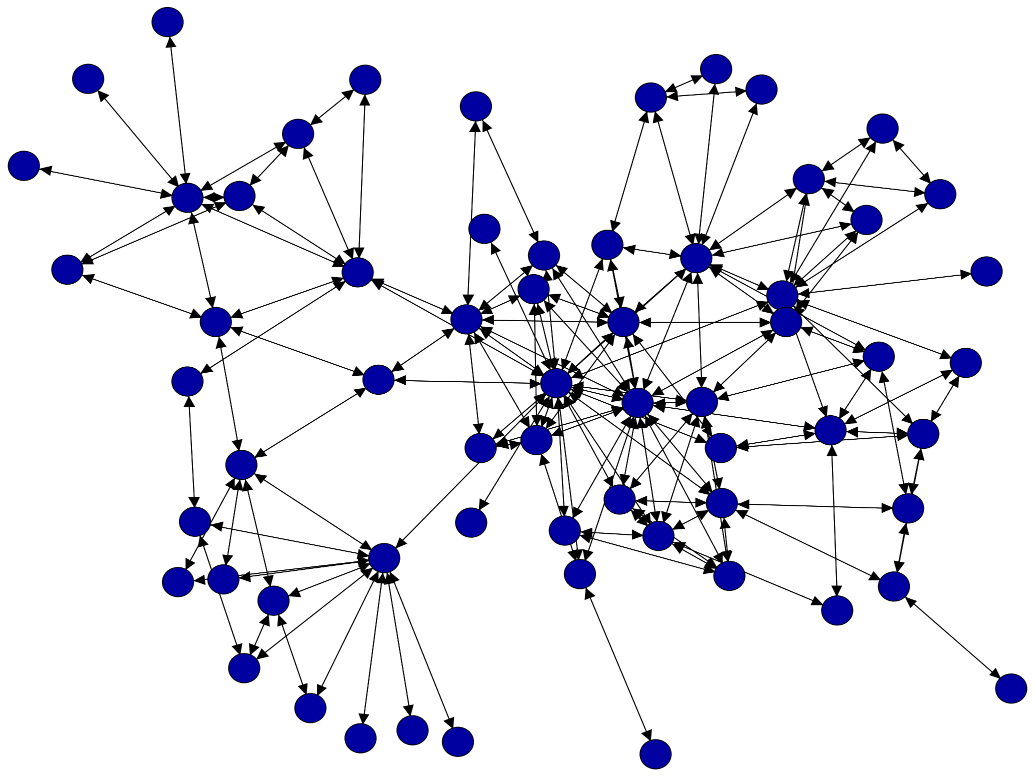
$$F = 1 - \frac{\sum_{i \neq j} r_{ij}}{n(n-1)}$$

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

Empirical Example #1
Disrupt Terrorist Network

DISRUPTION

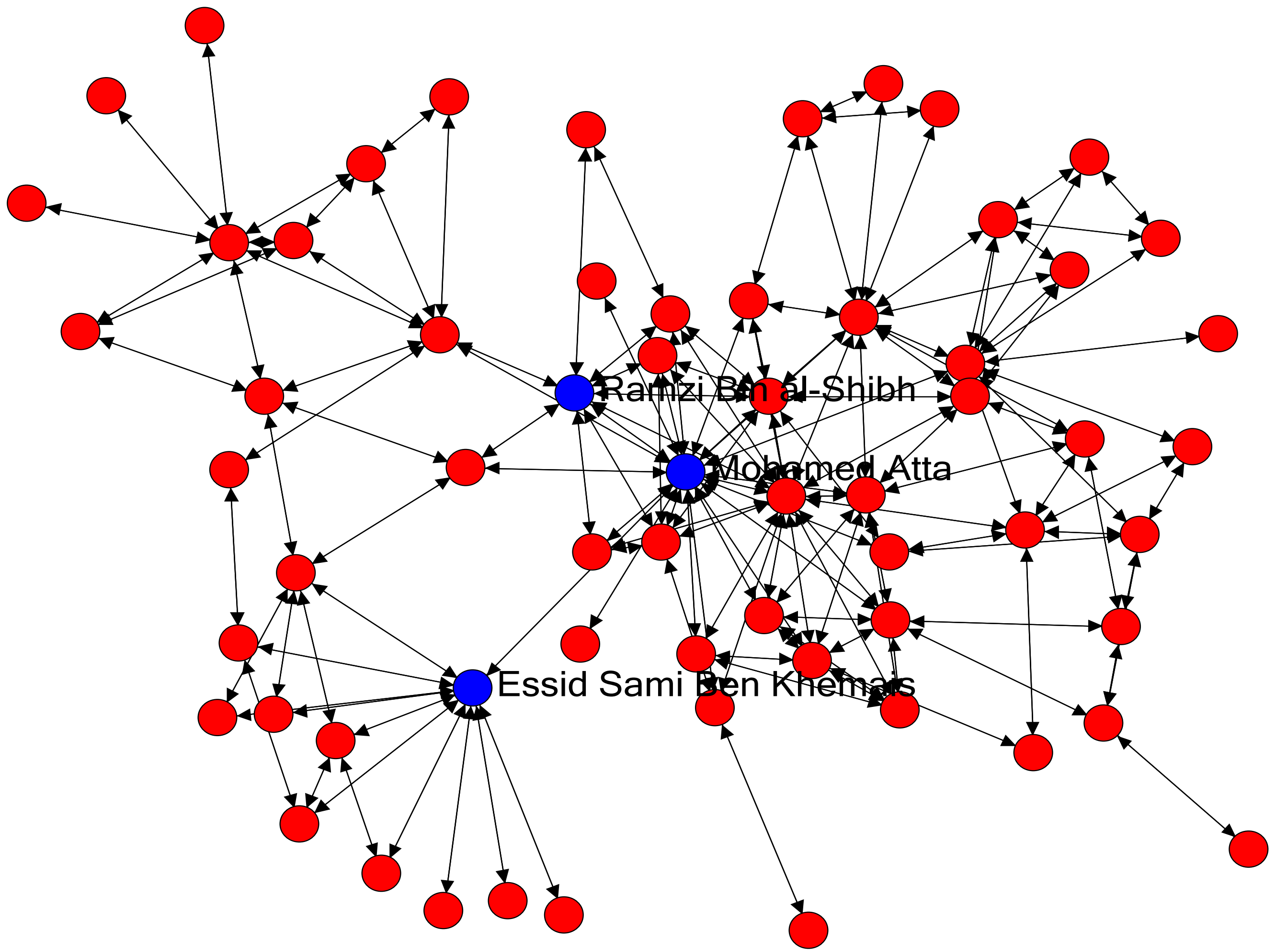
- Which three nodes should be isolated in order to maximally disrupt the network?



Data from: Krebs, V. 2002. Uncloaking terrorist networks.
First Monday 7(4): April. http://www.firstmonday.dk/issues/issue7_4/krebs/index.html

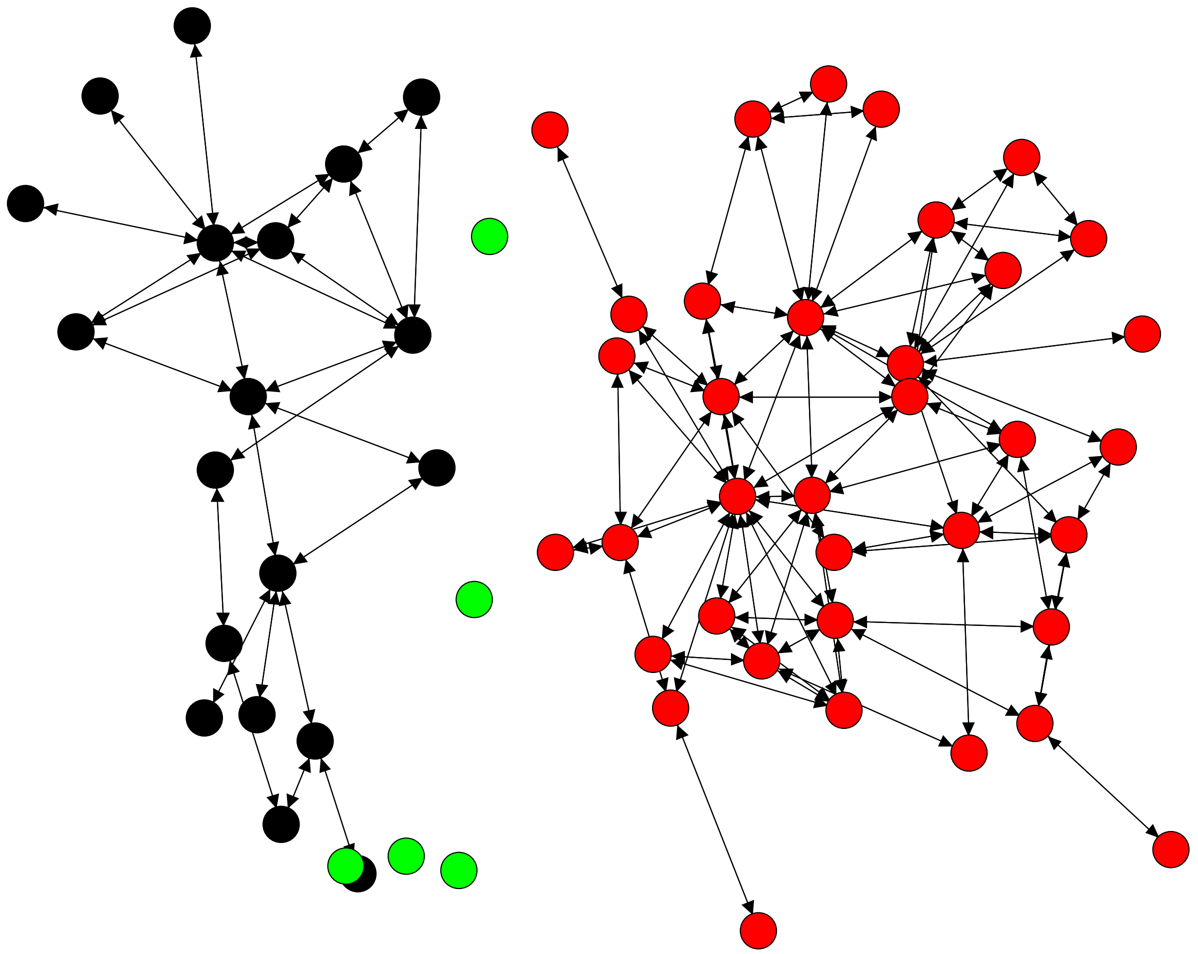
DISRUPTION

KeyPlayer Solution



DISRUPTION

KeyPlayer Solution
(key players removed)



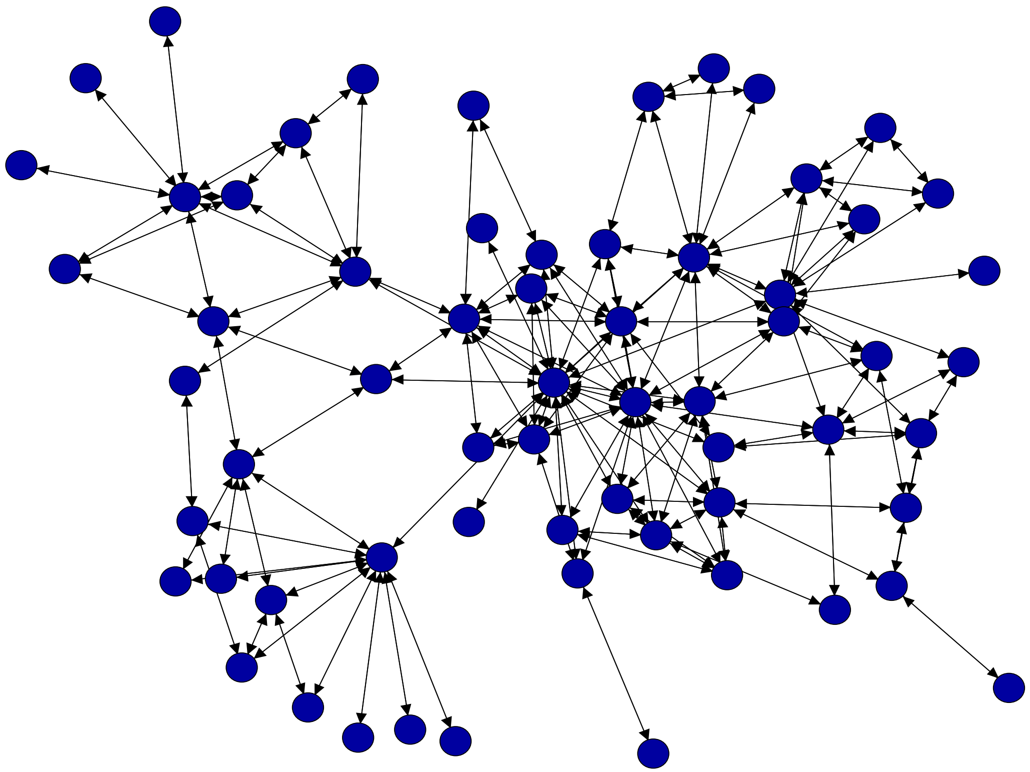
Why do we want to know who the key players are?

DISRUPT	We want to **remove them** – to maximally **disrupt** the network
ENHANCE	We want to **help** them – in order to make network as a whole **function better** (diffuse info; coordinate well)
INFLUENCE	We want to identify **key opinion** leaders – to **influence** the network
LEARN	We want to know who **is in the know** – so we can question or **surveil** them
REDIRECT	We want to remove/prune them – to **redirect flows** in the network toward our preferred players

INFLUENCE

Empirical Example #2
Influence Terrorist Network

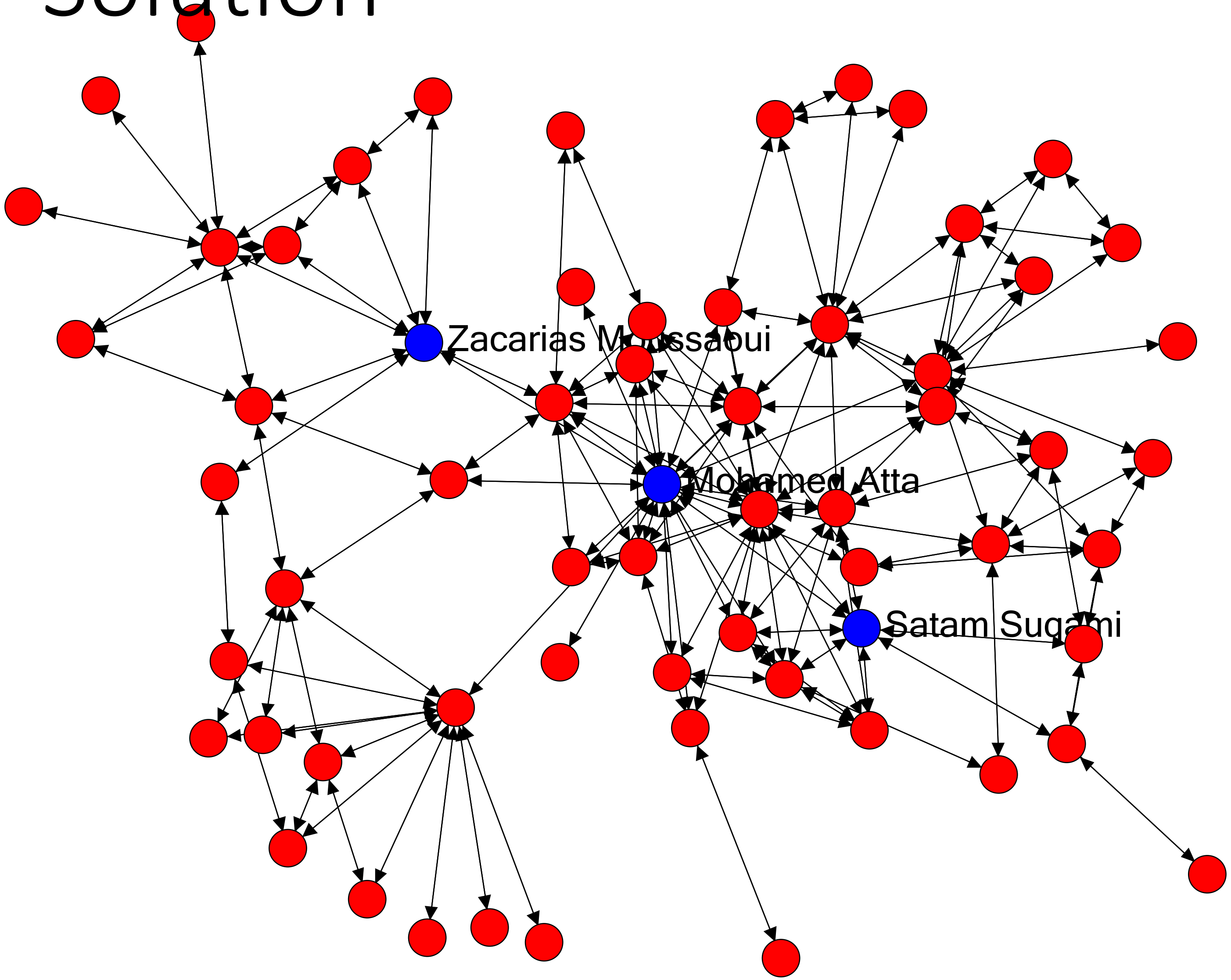
- Which three nodes should be selected in order to maximally influence the network by turning / planting information, etc.?



Data from: Krebs, V. 2002. Uncloaking terrorist networks.
First M w.firstmonday.dk/issues/issue7_4/krebs/index.html

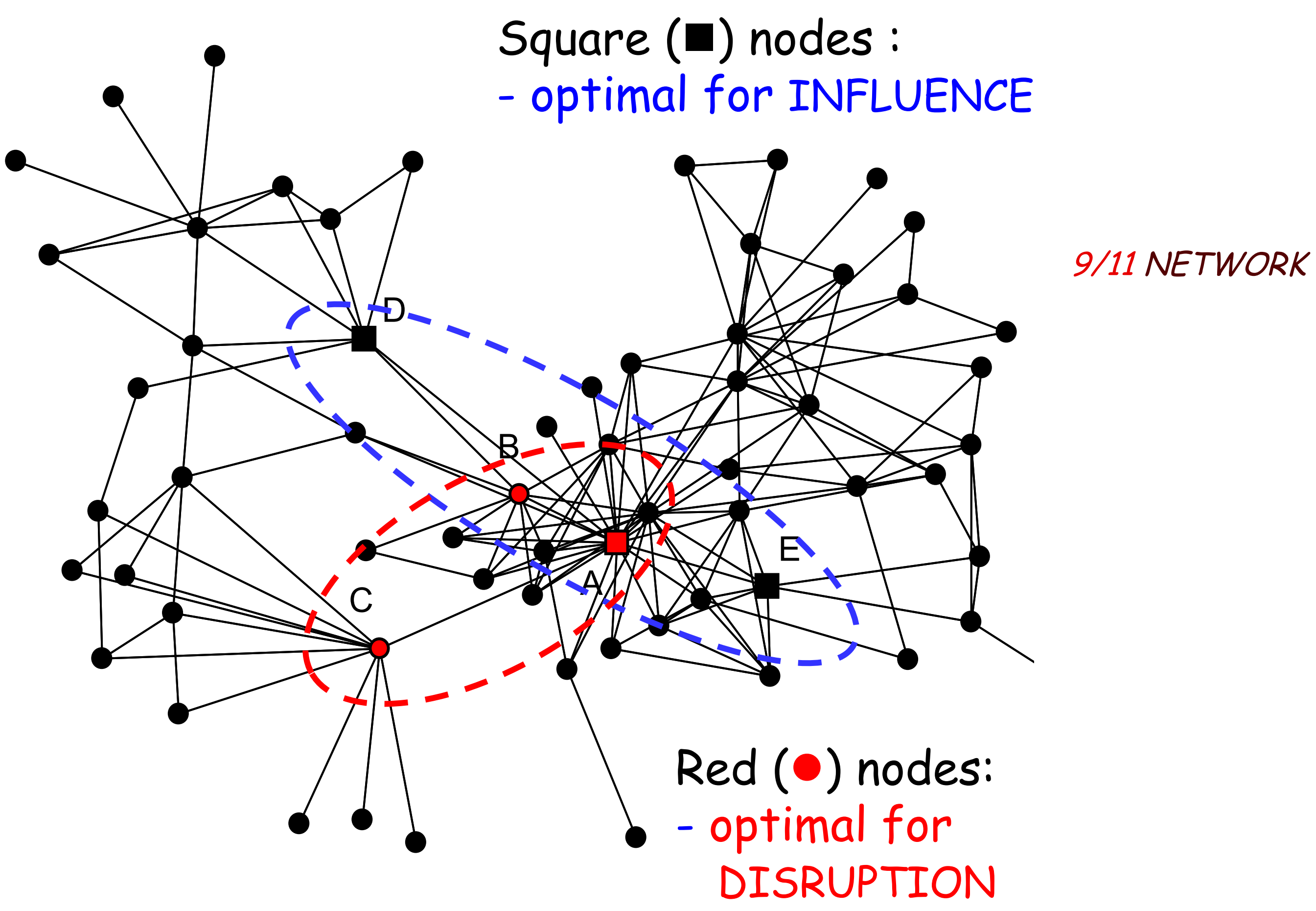
INFLUENCE

KeyPlayer Solution



Terrorist Network

- Red nodes identify optimal cho
  for DISRUPTION problem
  - Removing them splits network in
    7 components and yields
    fragmentation metric of 0.59
- Square nodes identify solution
  INFLUENCE problem
  - The best nodes to seed with
    disinformation

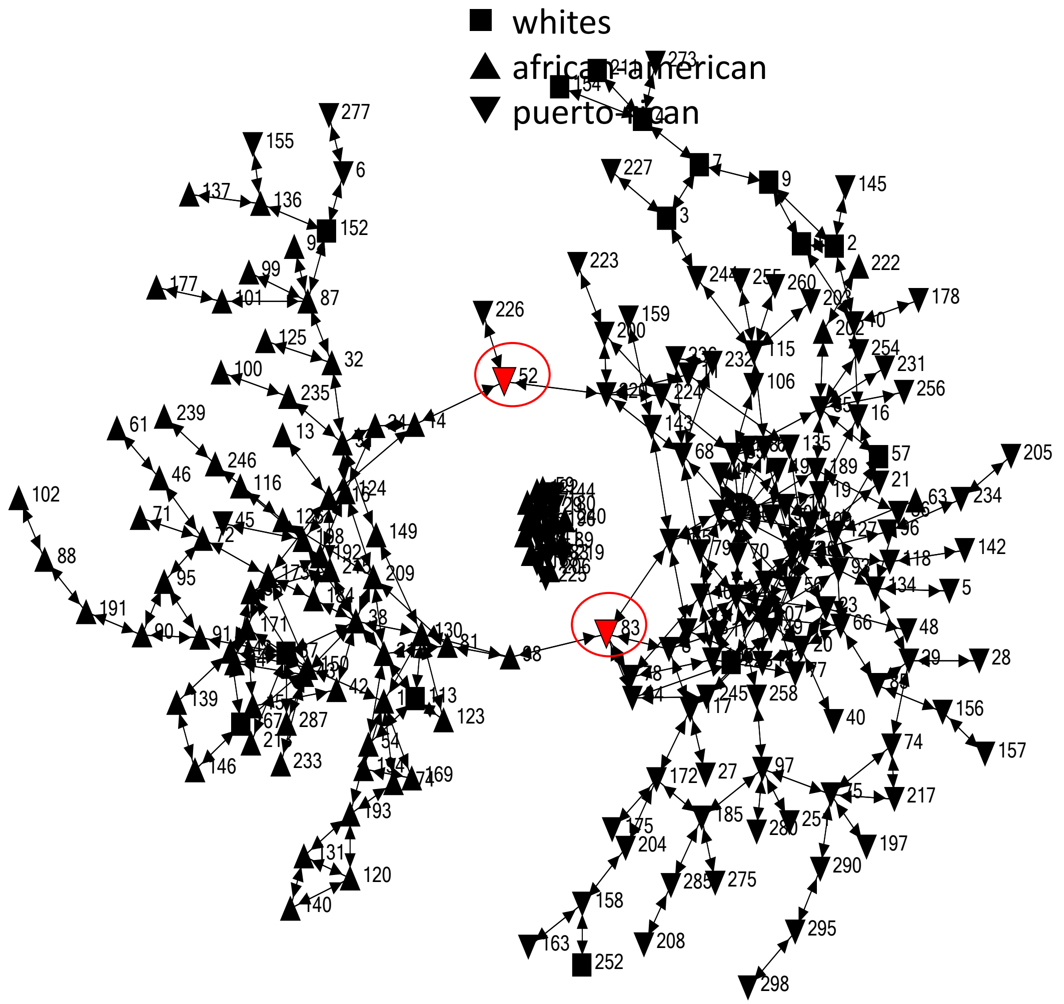


Data from: Krebs, V. 2002. Uncloaking terrorist networks.
First M stmonday.dk/issues/issue7_4/krebs/index.html

Disruption Example – health context

- Which two people should be isolate slow the spread of HIV?
  - KeyPlayer algorithm dc identifies the two red nodes

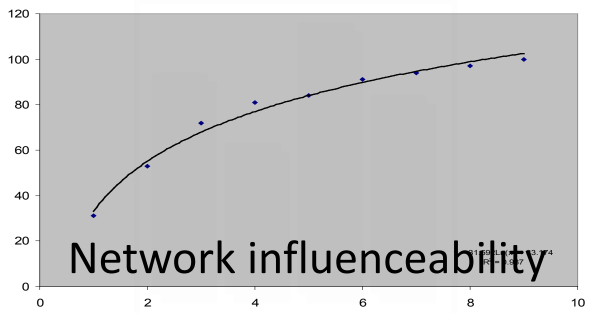
Friendship ties
among drug injectors
on streets of Hartford

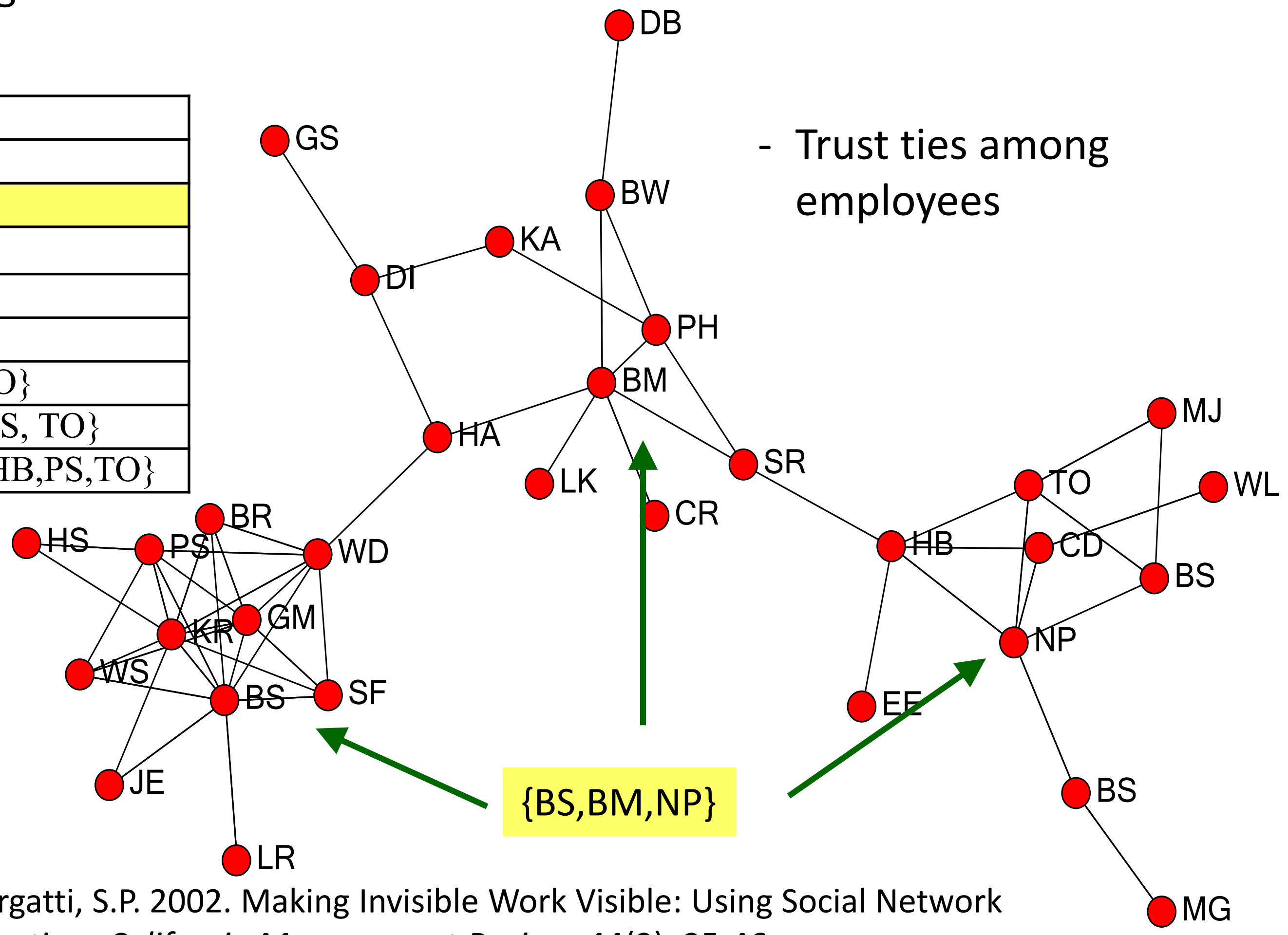


Influence Example – mgmt context

- Major change initiative is planned. Which small set of employees should we select for intensive indoctrination? in hopes they will diffuse positive attitude/knowledge to others

K	%	KP-Set
1	31	{KR}
2	53	{BM,BS}
3	72	{BM,BS,NP}
4	81	{BM,BS,DI,NP}
5	84	{BM,BS,DI,KR,NP}
6	91	{BM,BS,DI,HB,KR,TO}
7	94	{BM,BS,BS2,DI,HB,PS,TO}
8	97	{BM,BS,BS2,CD,DI,HB,PS, TO}
9	100	{BM,BS,BW,BS2,CD,DI,HB,PS,TO}





- Trust ties among employees

a from: Cross, R., Parker, A., & Borgatti, S.P. 2002. Making Invisible Work Visible: Using Social Network Analysis to Support Strategic Collaboration. *California Management Review*. 44(2): 25-46

Dyadic Cohesion

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

Average is density

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

Average is average distance

Minimum is line connectivity

Minimum is point connectivity

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 e.g., samesex → friendship Test models of tie formation | network change | selection	For each pair of nodes, predict similarity in choices as function of tie between them e.g., years of marriage → similar attitudes Test models of diffusion/contagion/influence
node	For each node, predict their centrality e.g., extraversion → number of friends Test models of social status attainment	For each node, predict success as a function of social ties e.g., friends in high places → business success Test models of social capital
group	For each group, predict the cohesion of network e.g., demographic similarity → density of ties	For each group, predict performance as a function of network structure 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 <u>random</u> 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

Friendship tie

$\approx$

	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

Age difference

$+$

	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

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

Friendship tie

$\approx$

	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

Age difference

$+$

	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

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

Friendship tie

$\approx$

	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

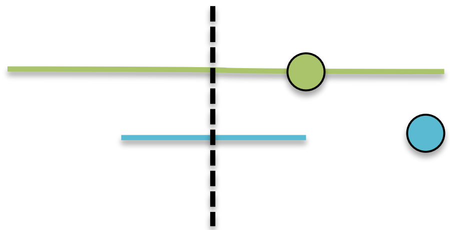
Age difference

$+$

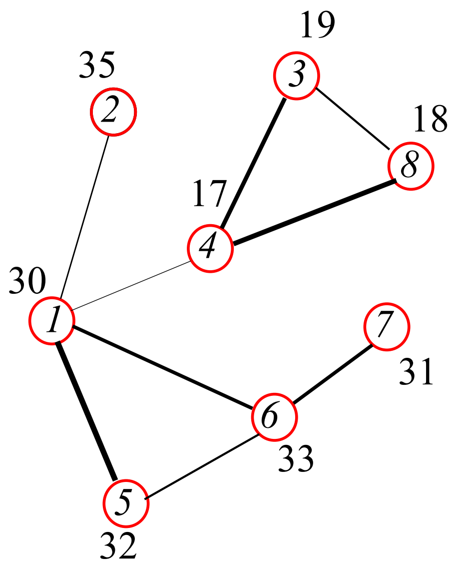
	A	B	C	D	E	F	G
A	0	1	0	2	1	0	0
B	1	0	3	5	1	4	2
C	0	3	0	4	5	8	10
D	2	5	4	0	0	3	2
E	1	1	3	0	0	2	2
F	0	4	2	3	3	0	1
G	0	2	1	2	2	1	0

education

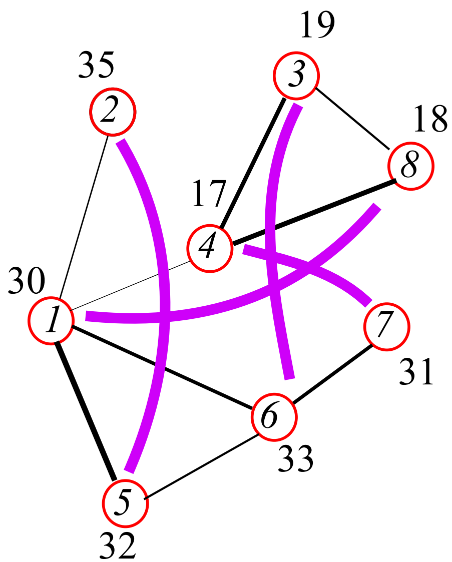
- Permutes dependent variables lots of time. Measure the sampling distribution of the coefficients.
- P-value is a proportion of times that the observation is Falling outside the sampling distribution.



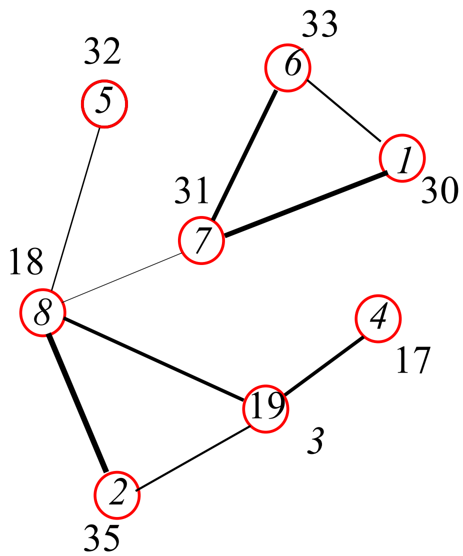
QAP process – graph representation



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)

Seek advice from = b0 + b1 Reports to + b2 Friendship

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

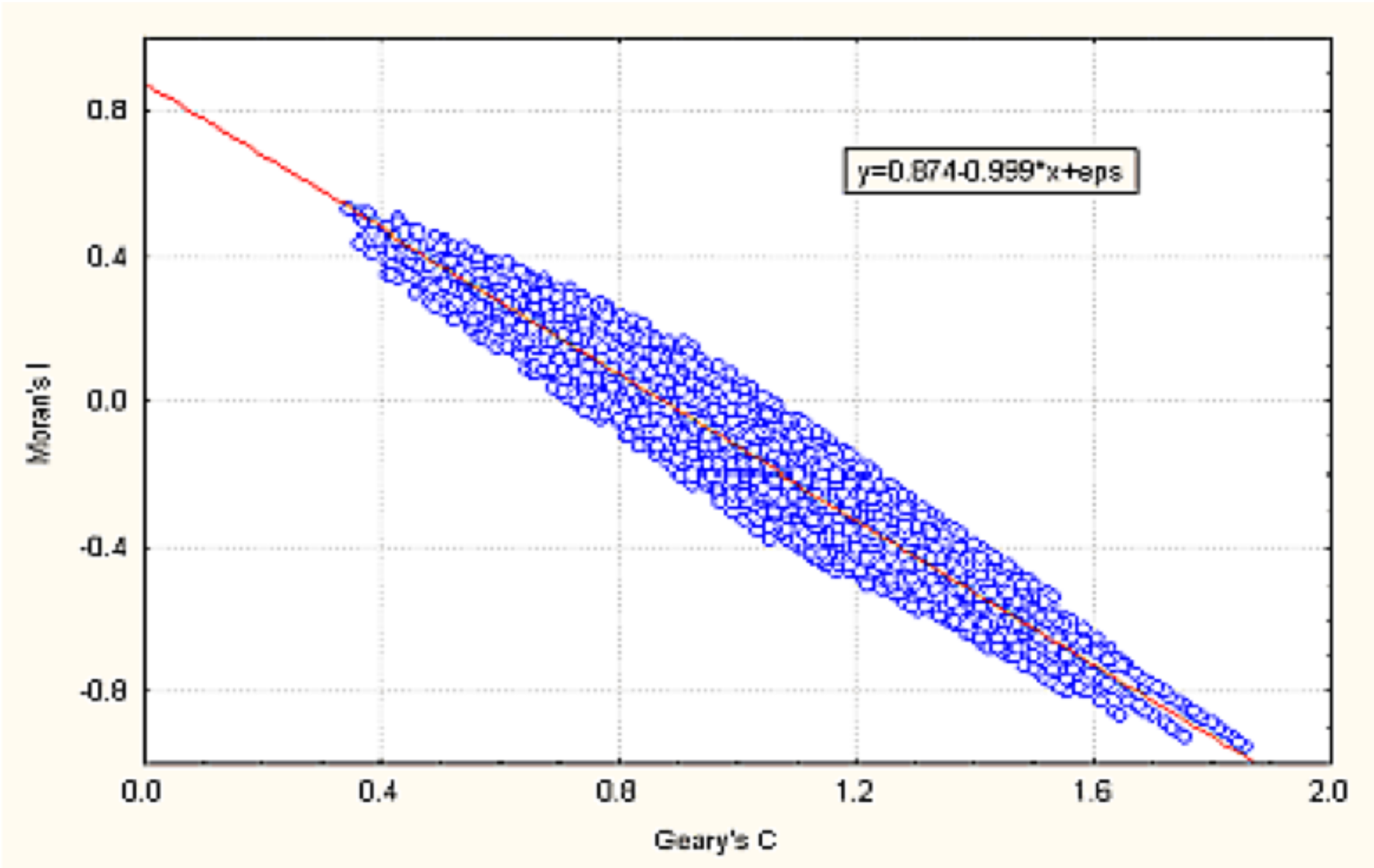
Continuous Autocorrelation

- 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

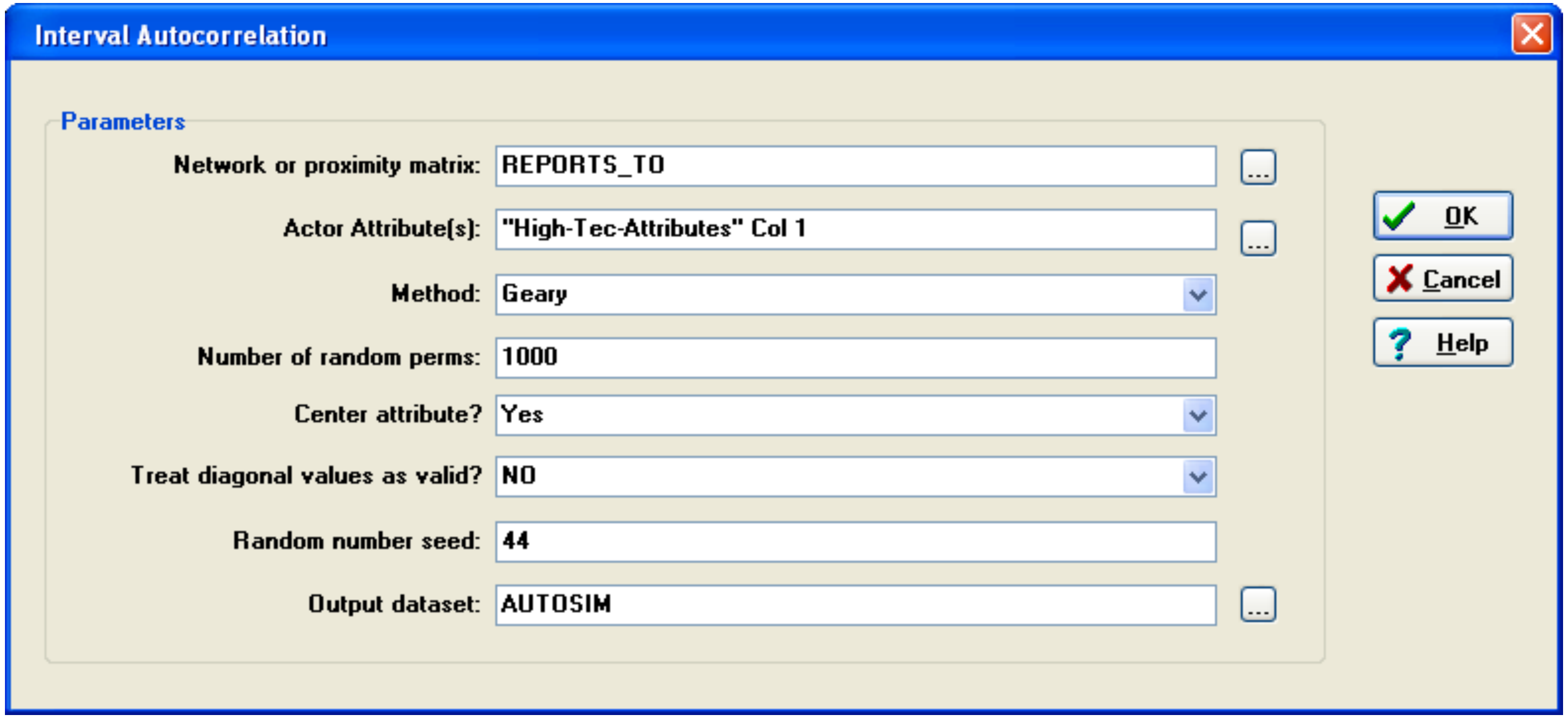
- Let $w_{ij} > 0$ indicate adjacency of nodes i and j, and X_i indicate the score of node i on attribute X (e.g., age)

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

- 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

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

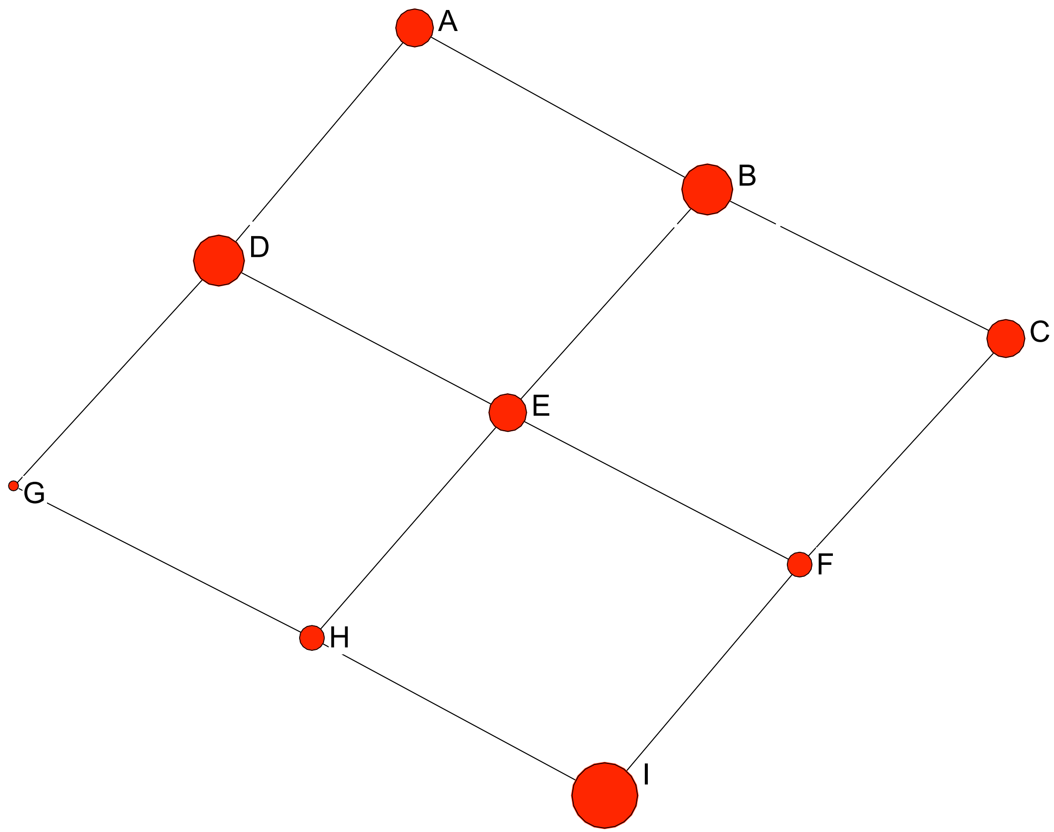
Moran's I

- 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)

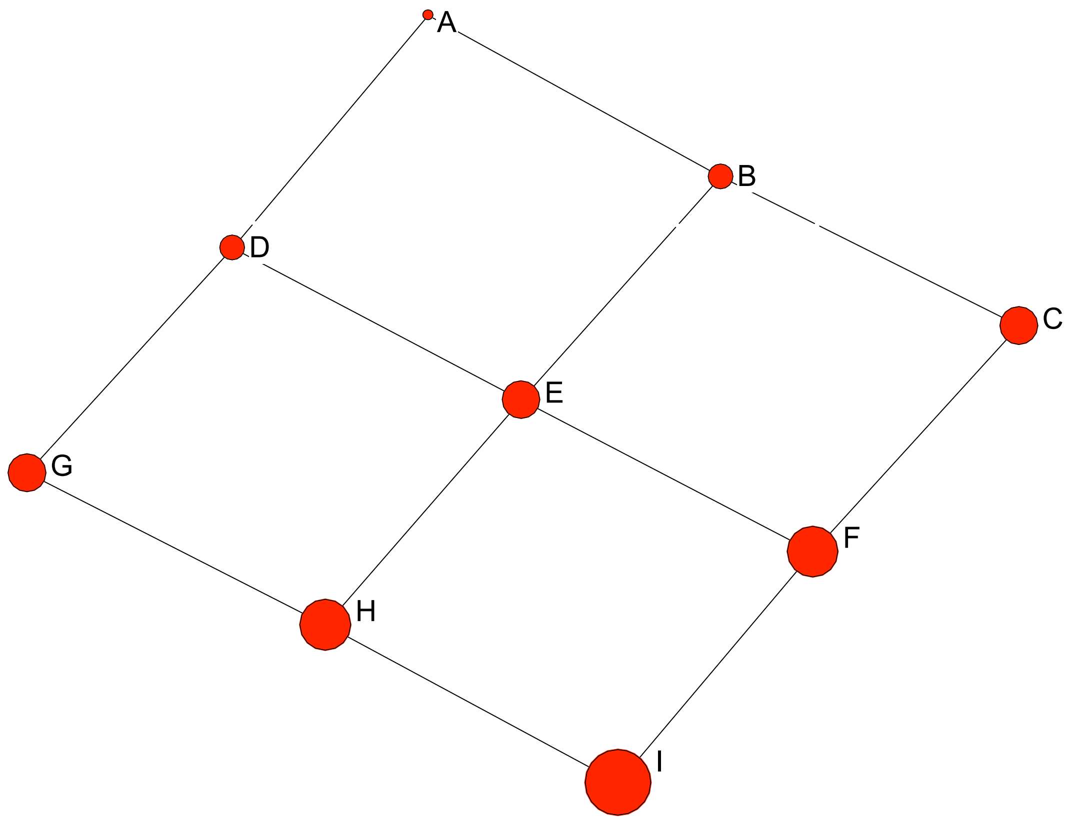


Node	Attrib
A	3
B	4
C	3
D	4
E	3
F	2
G	1
H	2
I	5

Moran's I: -0.250
Significance: 0.335

Positive Autocorrelation

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

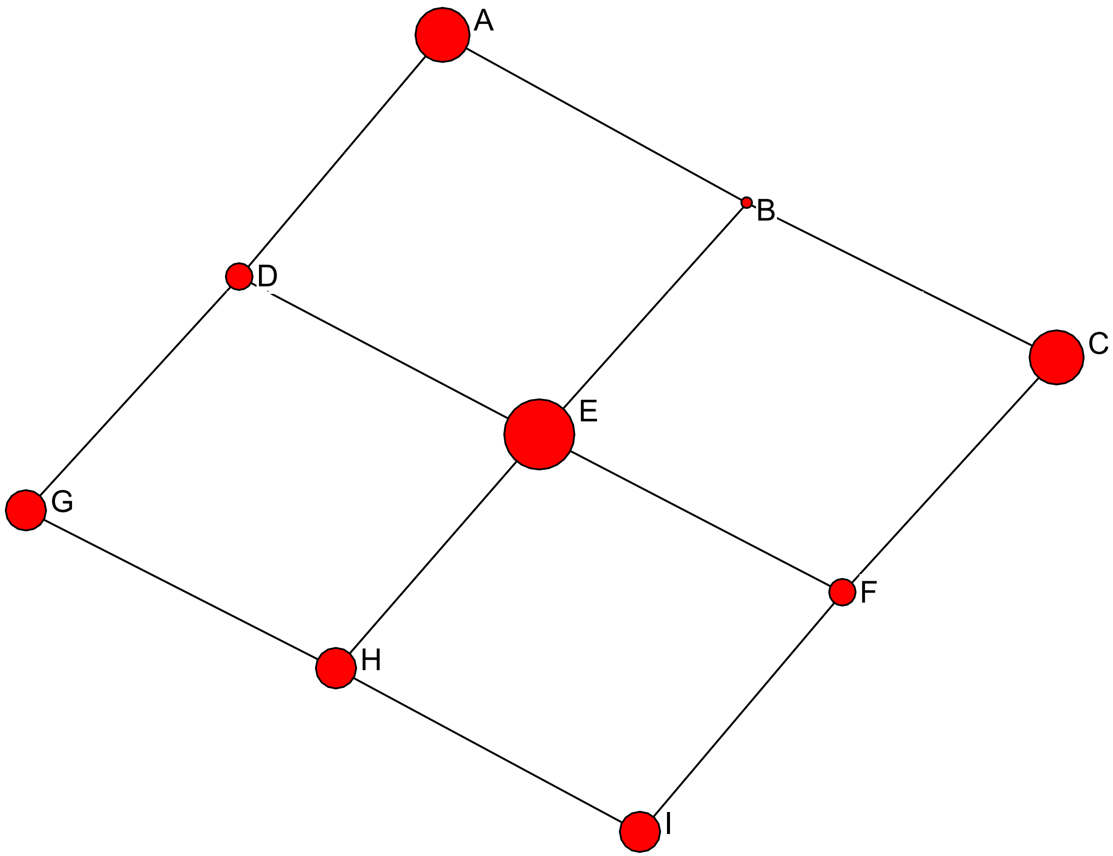


Node	Attrib
A	1
B	2
C	3
D	2
E	3
F	4
G	3
H.	4
I.	5

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

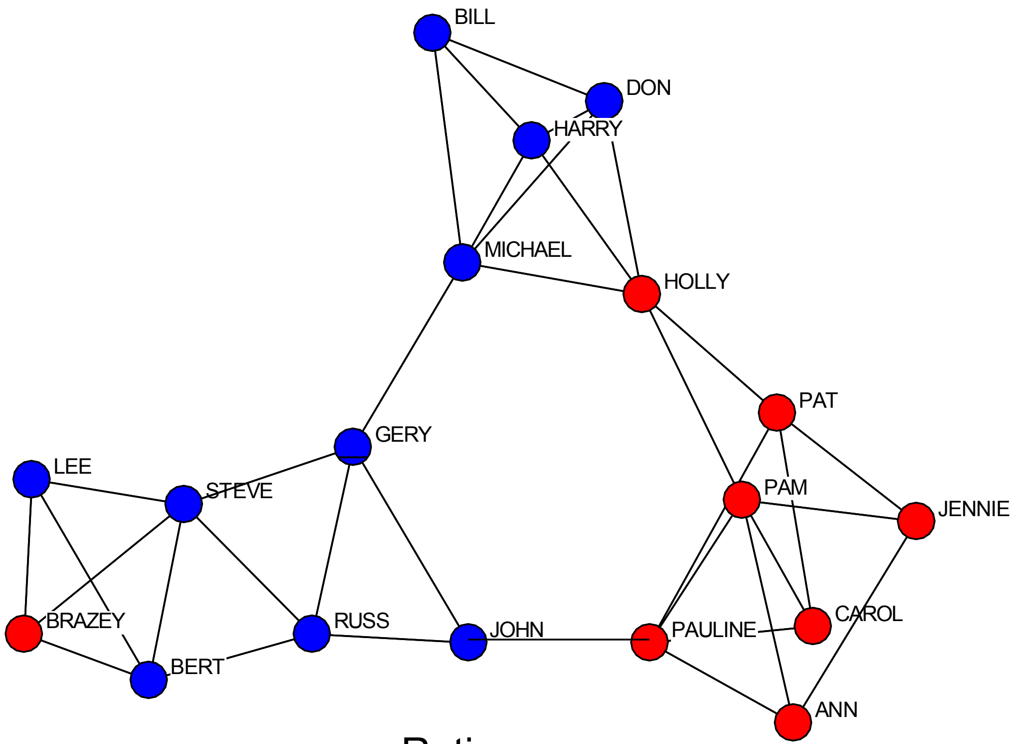
Interpreting Autocorrelation

- 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

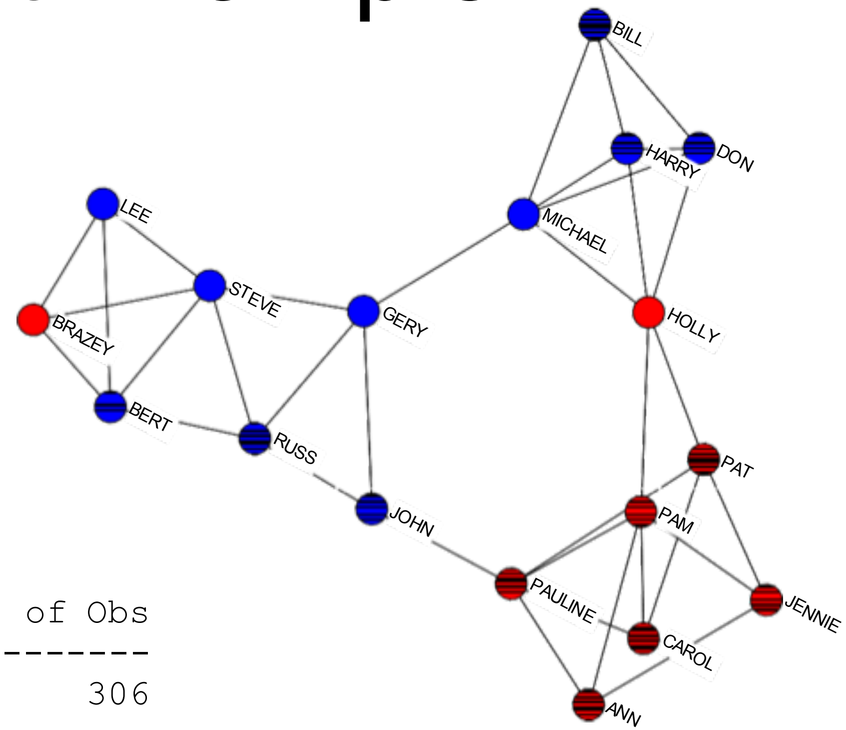
Expected

	Female	Male
Female	6.4	18.3
Male	18.3	10.3

Ratio

	Female	Male
Female	1.87	0.38
Male	0.38	1.55

Campnet Example



Density Table

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

MODEL FIT

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

REGRESSION COEFFICIENTS

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