Social Network Analysis

Data and Process

Robin Burke

DePaul University

Chicago, IL

Outline

- Questions
  - Homework I
- Social network analysis process
- Social network data
- Ethical issues
- Data preparation
- Sampling

Homework 1

- Questions

Teams Milestone

- Due next week
- Submit to D2L
  - Names of team members
  - 2 or 3 person teams
  - Everybody should do this (undirected network!)
- You can also submit
  - “Don’t have a team”
  - And I will form teams
- Use #project channel if you want to put together a team

Applying social network analysis

- Relational data
  - it has to make sense to treat the data as a network
  - does it matter that there is a path from one individual to another?
- Relational questions
  - questions must go beyond what you can get from a table
    - counts, relative proportions, etc.

Relational questions

- Who are the important individuals in the network?
  - (What does it mean to be important?)
- What are the sub-groups in the network?
  - (How are sub-groups defined?)
- What roles do individuals play in the network?
  - (What roles matter? How to detect?)
- What influences led to the formation of this network?
  - (What kinds of influences are we talking about?)

Stages of social network analysis

- Network definition / collection
- Manipulation
- Calculation
- Visualization

Network definition / collection

- The “true” social network is the whole planet
  - “no man is an island”
- Always have to set boundaries
- Cannot apply sampling in the same way
  - as traditional social science research
- Collecting social networks can also be difficult
  - the Internet and the social web has made it a lot easier

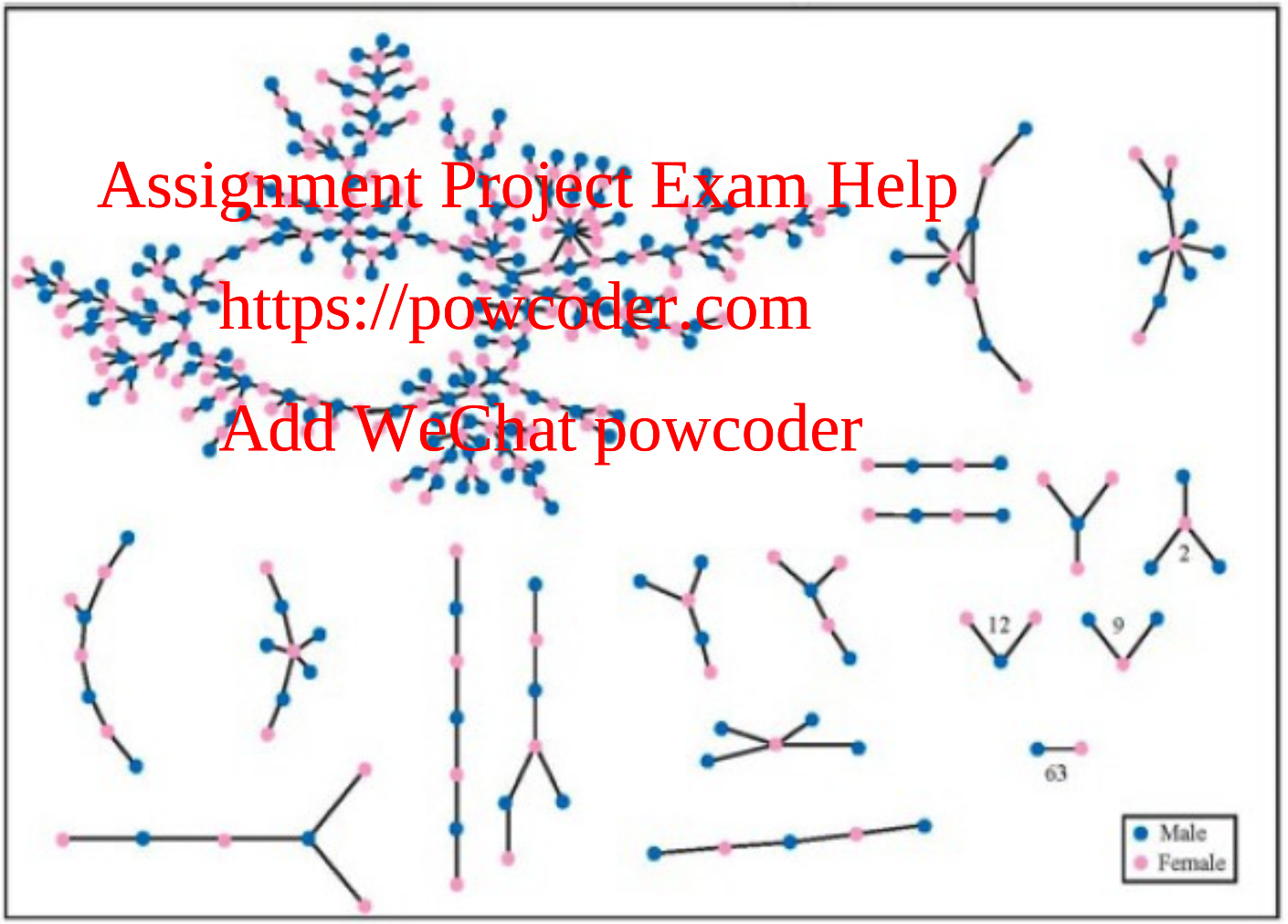
Manipulation

- Recovering the network may require computation
  - individual emails with “from” and “to” fields
- Data cleaning is important
  - Data has to be filtered, merged, normalized, etc.

Calculation

- We'll spend a lot of time on this
- What can we measure about a network?
- What does it make sense to measure about a network?

Visualization



Visualization

- Network data is inherently complex
  - visualization is often the best way to make sense of it
- But you can also make bad visualizations
  - we'll talk about how to avoid that

Stages of social network analysis

- Network definition / collection
- Manipulation
- Calculation
- Visualization

Network Definition

- Basic questions
  - what are nodes?
    - usually people, but not always
  - what are edges?
    - usually relationships, but how defined?

Boundaries

- Very important question
  - what is the set of individuals under consideration?
    - where do we stop?
- Part of any data description

Network relations

- Not every relation is an edge
- For example
  - I could create an age-based bipartite network
    - link people by their ages
- Why not?
  - age doesn't have "path semantics"

Problems with the age network

- People only have one age
  - Connected only to people of the same age
  - This is the same information you could get by creating a table of people vs age
- More generally
  - there is no utility to paths in this network
  - true of some multiplex relations as well
    - citizenship

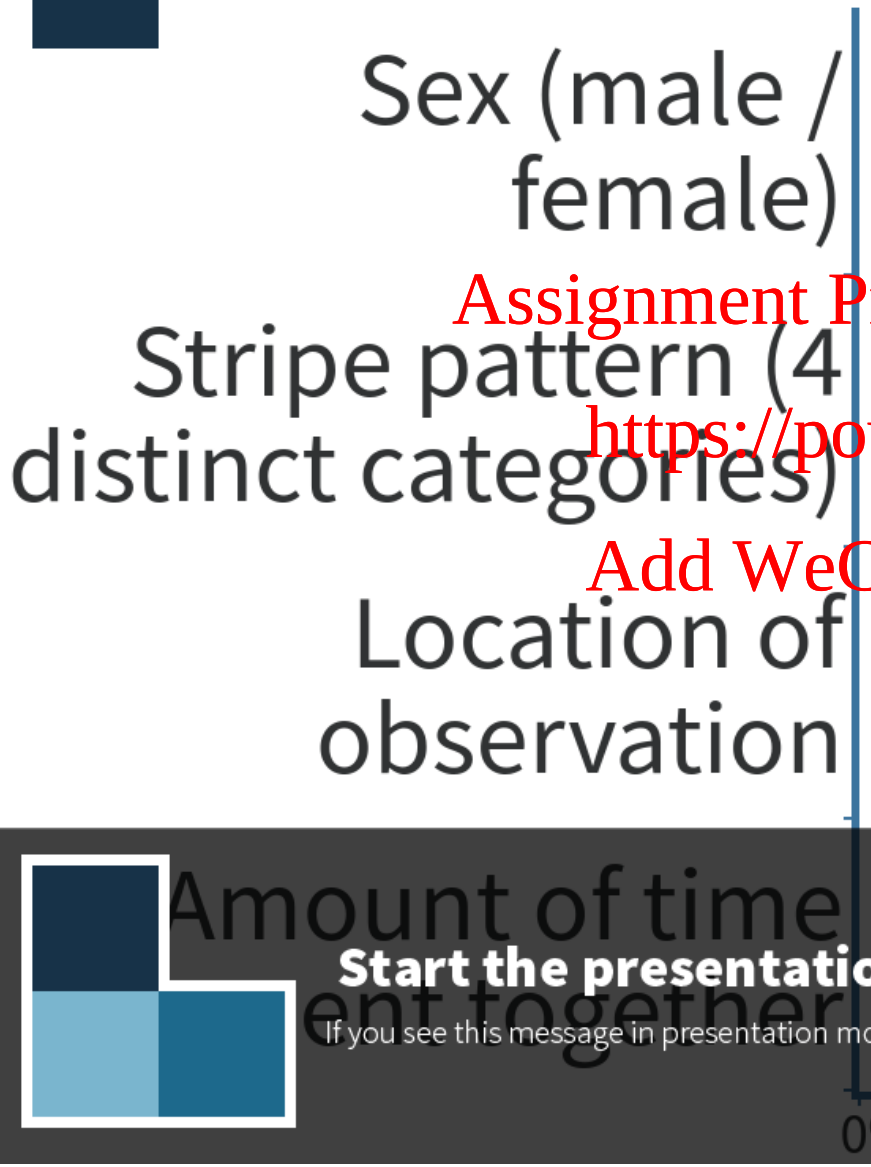
Network relations / edges

- Will be multiple
  - more than one connection for each node
- Will have path semantics
  - paths of arbitrary length have meaning

Measuring relations

- Binary
  - tie exists or not
- Heterogeneous
  - what kind of tie?
- Ranked ordinal
  - "close" vs "distant"
  - even "liked" vs "disliked"
- Interval measures
  - harder to get from people
  - easier to get from machines
    - e.g. # of interactions

observing zebras in the wild. What would be
relation for building edges?



Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Gathering data

- Go out and get the data you've defined
- Often the network must be redefined
  - issues arise while gathering it
- If you have a network someone else has gathered
  - very important to know how they defined the network when collecting it

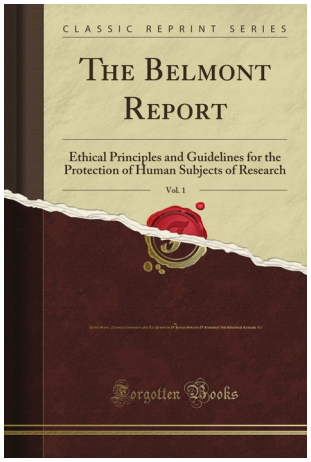
Ethics

- Social network data is often easy to collect
  - public sites / APIs
- What are the ethical considerations?

Key principle

- Social network data is data about people

Belmont Principles



- Belmont Report, 1978
- Autonomy/Respect for persons
  - respect for individual autonomy, and particularly protection of persons with diminished autonomy
- Beneficence and Nonmaleficence
  - maximize benefits and minimize harms
- Justice
  - benefits and burdens should be justly divided

Association of Internet Researchers

- The more vulnerable the community/participants, the greater the obligation of the researcher
- Harm is contextually defined, hence ethical principles are inductively understood, in a context-dependent manner
- Data comes from people. You work with data generated by people, thus you work with people. Personal information involves a person in the end, even if the relationship is not always obvious.
- Balance the rights of subjects with benefits of research
- Ethical issues arise at every step, from planning through research through dissemination
- Ethical decision making is a deliberative process, it is best to consult widely
- https://aoir.org/

Four Research Types

I. User awareness and manipulation

Lab-based user study

II. Awareness without manipulation

User diaries / focus groups

Many corporate settings

III. No awareness with manipulation

A/B testing of new feature

IV. No awareness, no manipulation

most observational studies

This class

What is the harm?

- Data can be used to infer attributes of individuals
  - Including attributes they might want to keep private
  - Sexual orientation, health status, etc.
- Social networks are intensely personal
  - Who you interact with, how much?
  - How long have you know a given person?
  - Comments may reveal political preferences, income level, etc.

Primary obligation

- Remove personally identifying information (PII)
  - If the data is in any way non-public
  - Example:
    - You build a network from your own Facebook network
  - If you're drawing conclusions of a sensitive nature, even if the data is public
  - Example:
    - Detecting sexual orientation from Twitter posts
- Treat any link between identifiers and data as confidential

PII

- Is not just names, addresses and identifiers
  - Depends on the size and homogeneity of the group
- Example
  - In a smaller group, name of high school and year of graduation may uniquely identify individuals

Ethics = balance

- Balancing the concerns of individuals and those of society
- Specific scenarios are important
- Issues
  - benefit (who gains?)
  - cost (who loses?)
  - expectation

For our purposes

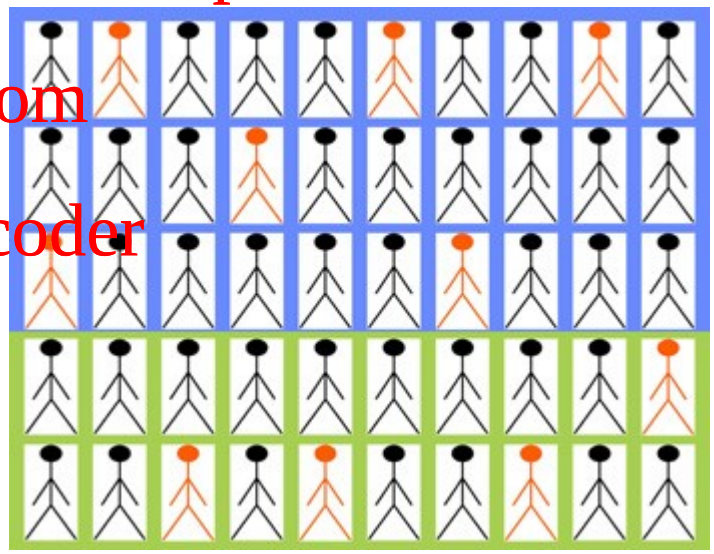
- If you're collecting data
  - think about
    - what kind of action resulted in that data
      - public or private?
    - what is the reasonable "expectation of privacy"?
  - consider possible harm
    - what if someone reading your final report knew an individual involved?

Example

- Dining social network
- Potential benefit?
- Potential harm?
- Mitigation strategy?

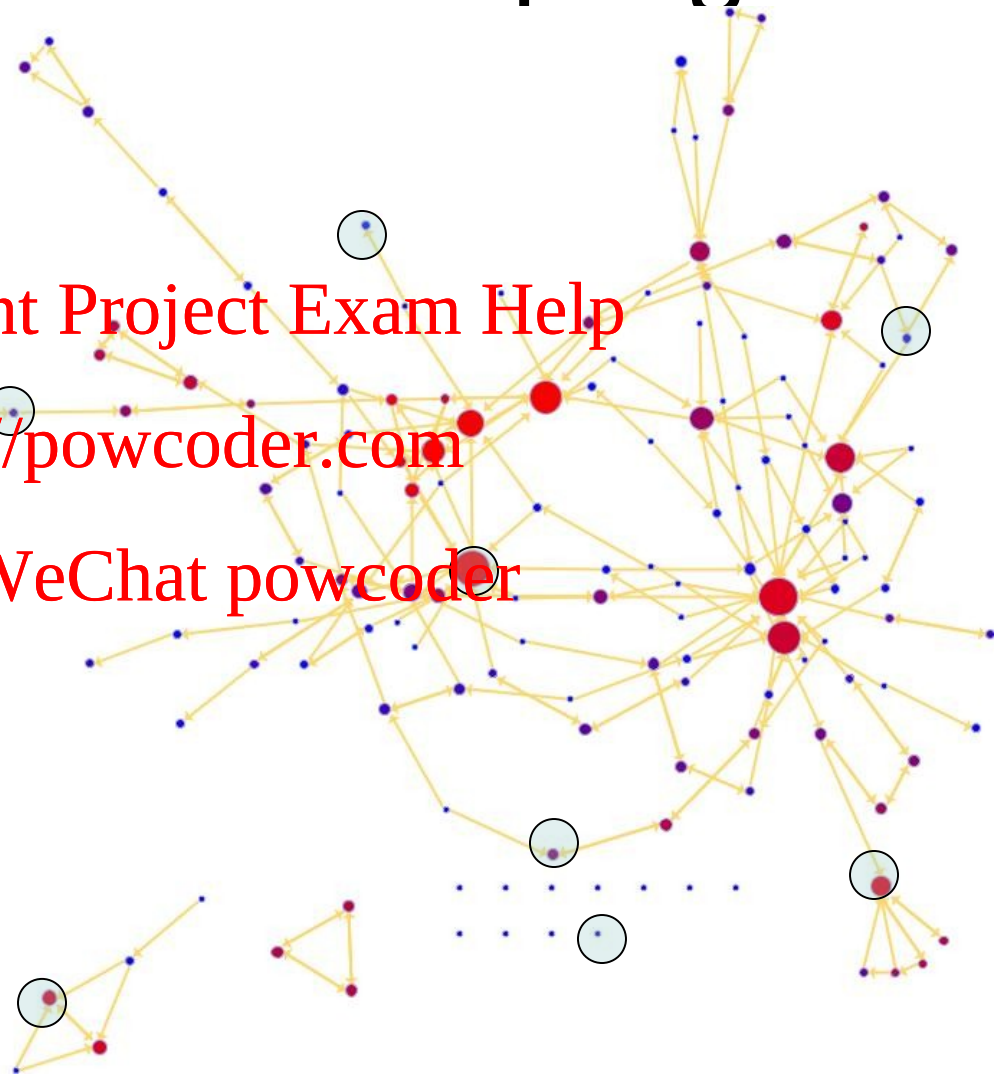
Sampling

- Social science research often depends on sampling a population
  - think election polls
- Identify individuals at random from a population
  - test their properties
  - extrapolate to the population as whole



Social network sampling

- Random doesn't work so well
- Have to make sure that the local properties of the network are captured
- Becomes a complex question



Full network

- You are interested in a known set of individuals
  - collect all the connections between these individuals
  - (not sampling)
- Example
  - all students in this class
  - all passengers on the Titanic
- Not always possible to do this

Reset Boundaries

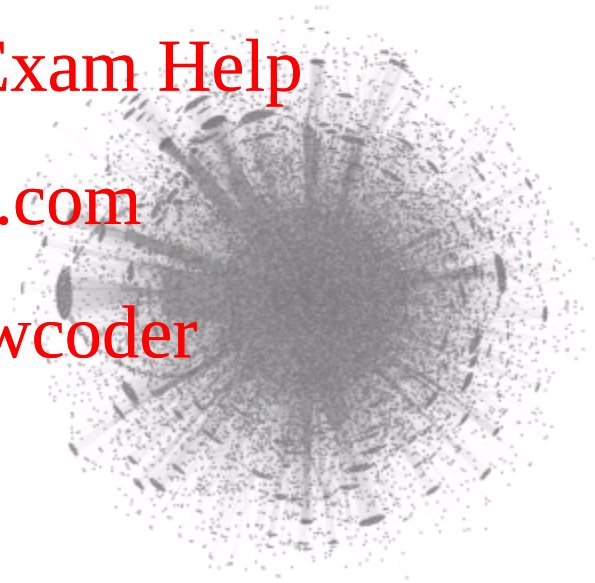
- If the data is
  - too big, or
  - can't be accessed in raw form
- You might need to establish new boundaries
  - collect within those
- Example
  - all users who tweeted using a particular hashtag in a given week
  - all students in a certain degree program

Sampling

- With online data (social networks)
  - you often don't have access to raw data
  - limited access via API
- Must be aware of the effect of not having all the data

Enron email network

- All people who exchanged at least 10 messages
- Pretty hard to make sense of this

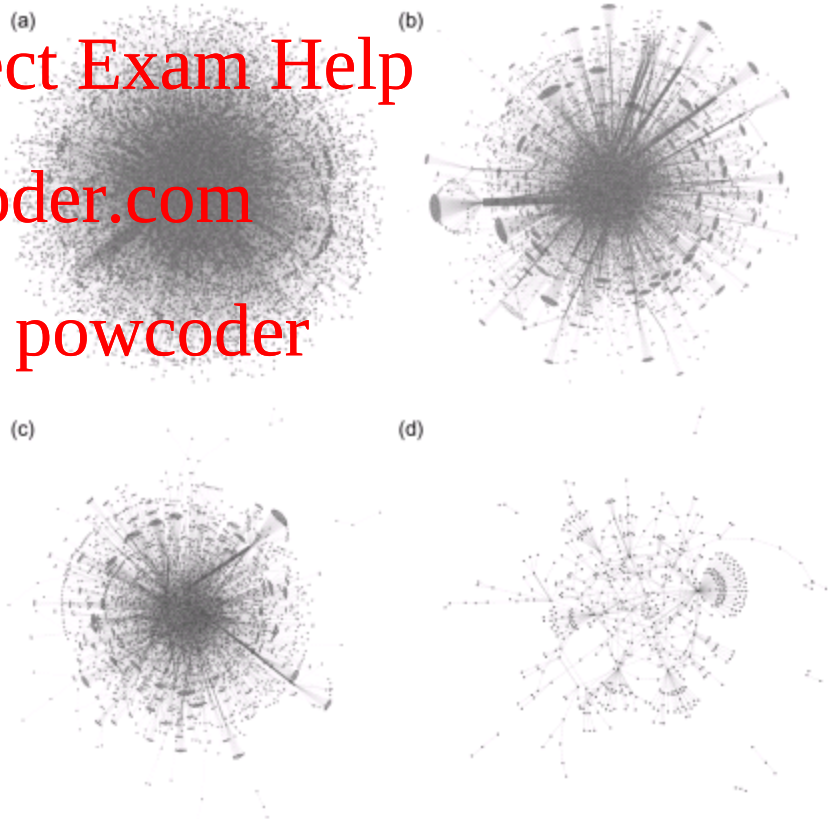


Random Sampling

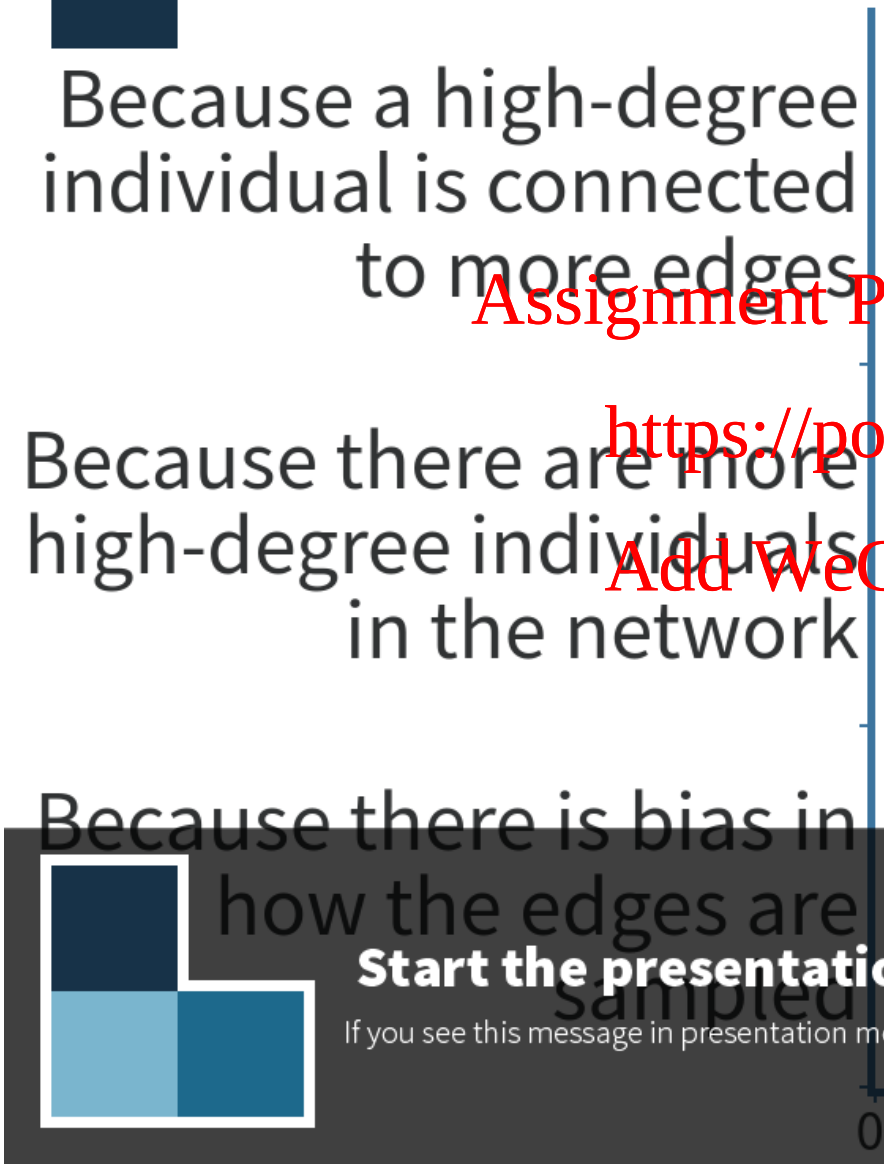
- Randomly select a certain percentage of **nodes** and keep all edges between them
  OR
  Randomly select a certain percentage of **edges** and keep all nodes that are mentioned.
- Problems
  - Edge sampling biased toward high degree nodes
  - Node sampling loses structural characteristics
- Benefits
  - Easy
  - Node sampling keeps some network statistical features

Enron network (random edges)

- Edge sampling
  - 50%, 25%, 10%, 1%
- Some structural properties preserved
- Biased towards high-degree individuals



a network formed by edge sampling more li
high-degree individuals than low-degree indi

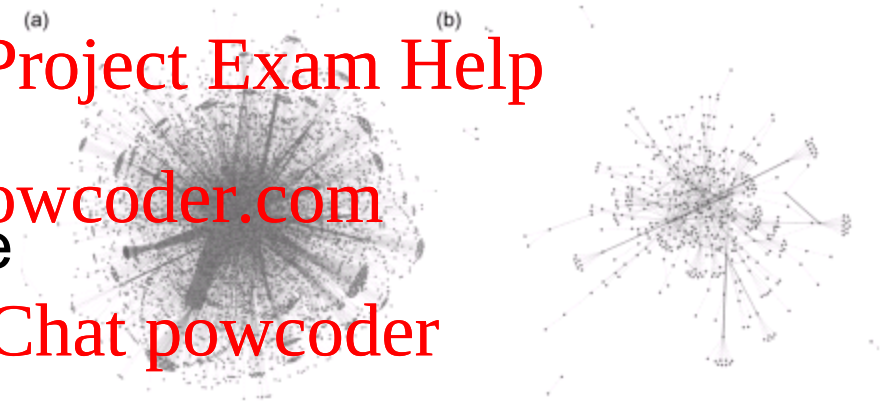


Start the presentation to activate live content

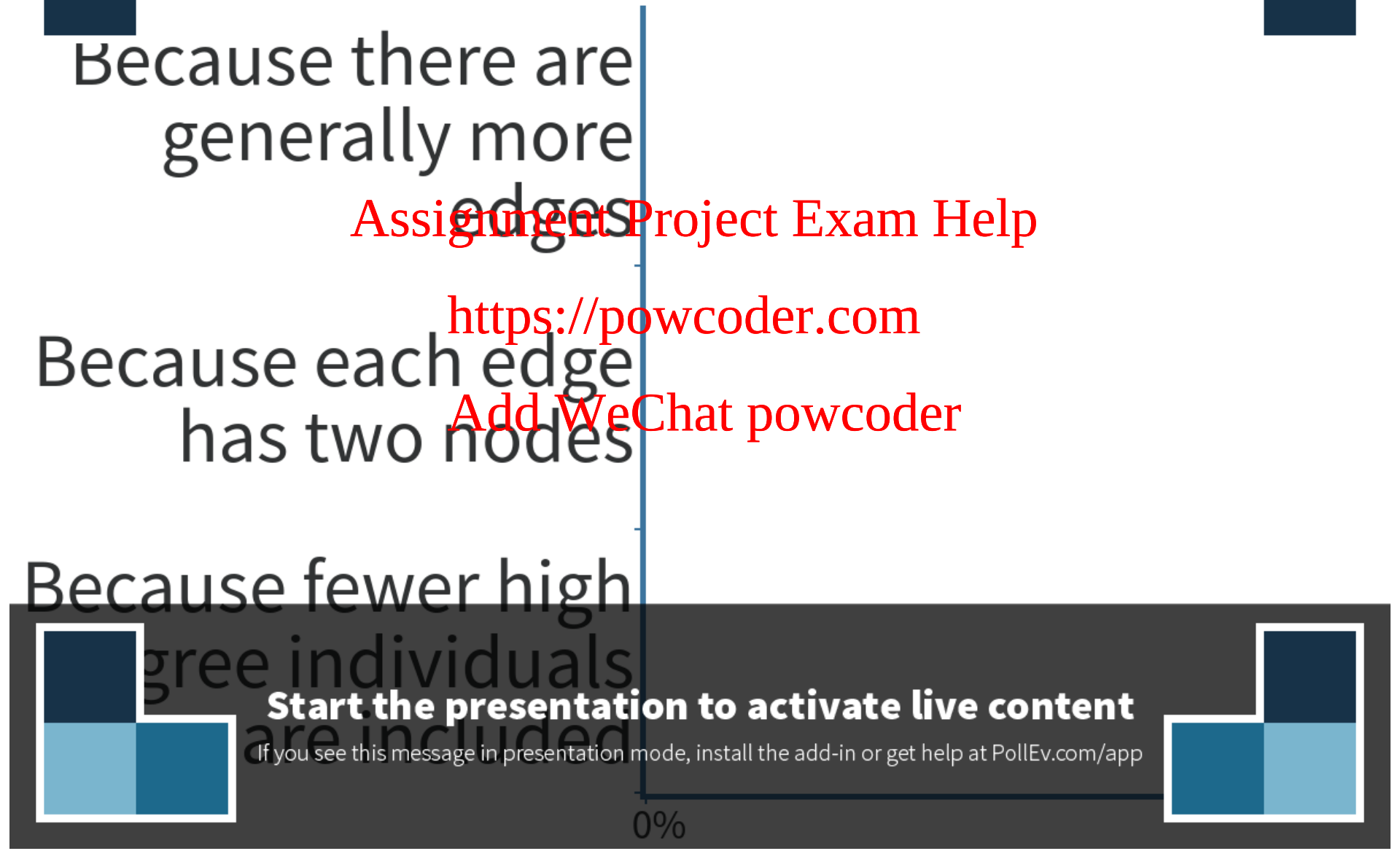
If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Enron network (random nodes)

- Node sampling
  - 50%; 10%
- Much less dense than edge sampling
- Less bias
- Node metrics can still be computed



s a given degree of node sampling (say 10%)
er network than edge sampling at the same



Survey methods

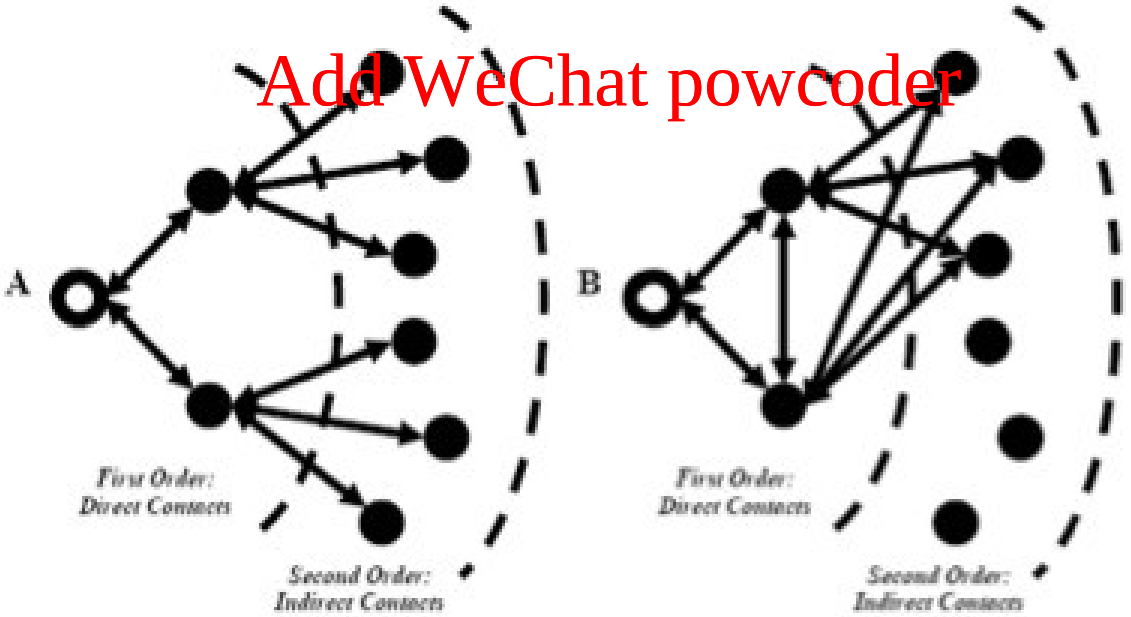
- In the social sciences
  - standard methodology is to survey individuals
    - ask them to name all of their ties
    - "who do you ask for advice?"
  - or collect observations of who interacts with whom
    - hard to ensure completeness
    - for some domains, this is all there is: animal social networks
- In online social networks
  - more likely to look at observed on-line behavior
  - but there may be unobserved off-line activity

Snowball Sampling

- When working with a large network, choose a starting node
- Get that node, its connections, their connections, and so on until the network is the right size for analysis
- Problems
  - Biased toward the part of the network sampled
  - May miss large-scale features
  - Cannot find isolates
- Benefits: Easy to do, common
- Might have no other options

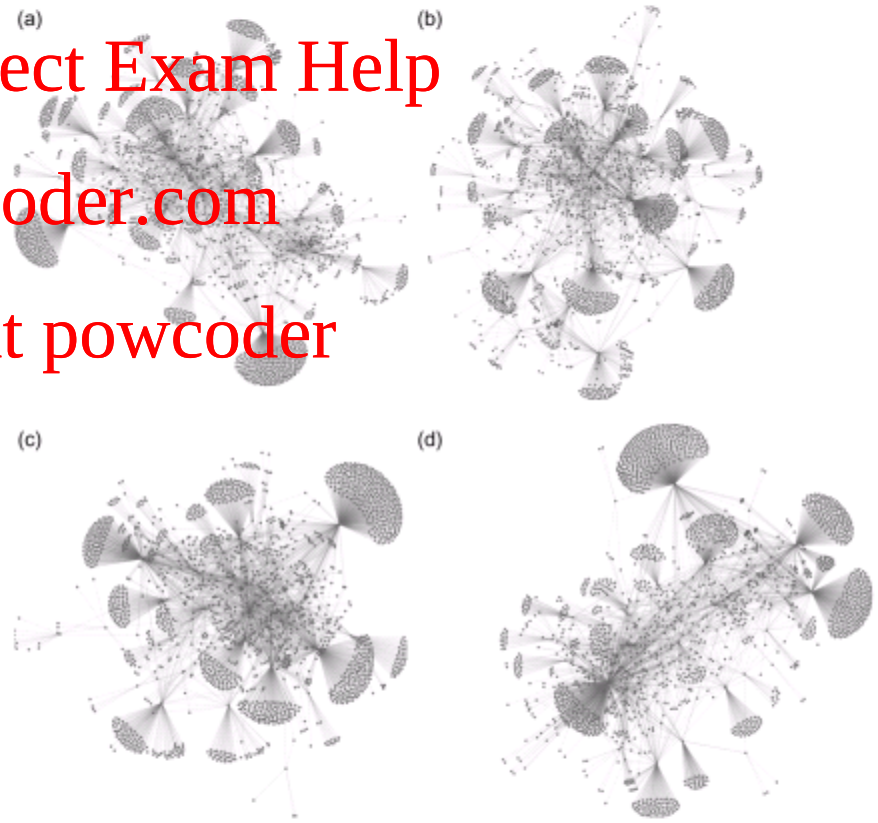
Sampling contacts

- Most prominent contacts are most likely to be shared
  - so a short contact list may not show the full network
- May need a longer list of contacts
  - sample from those



Enron network (snowball)

- Snowball samples of size 4
- Different starting points
- Note the “fans” at the edges
- Cannot measure network metrics reliably

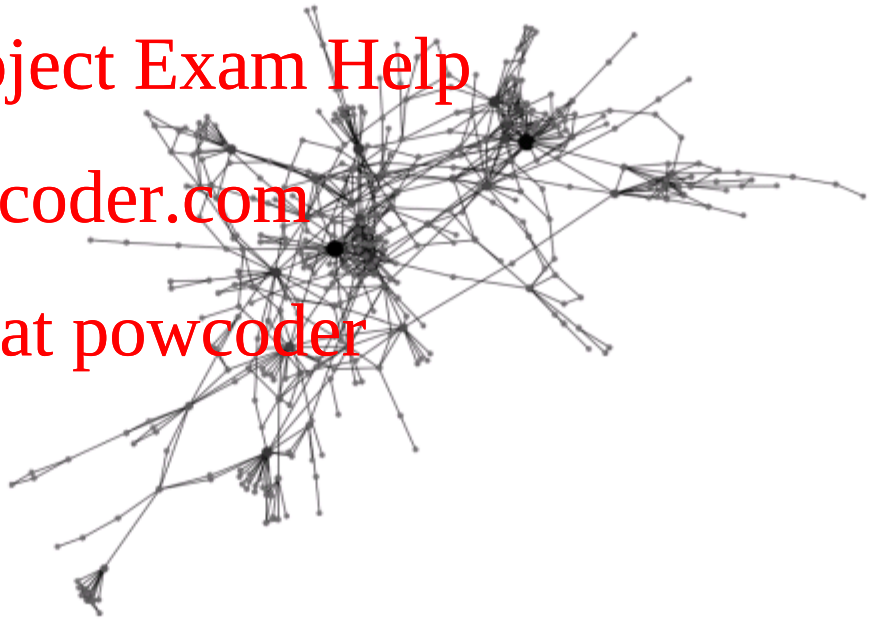


Filtering

- You can filter the network
  - keeping particular nodes or edges of interest
- Usually
  - interested in the most active users
  - the strongest ties
  - 80 / 20 rule

Enron network (filtered)

- Keep edges with 100 emails
- Keep edges if they account for at least 10% of total email output for a user
- Discard nodes of degree 1
- Keep only the giant component



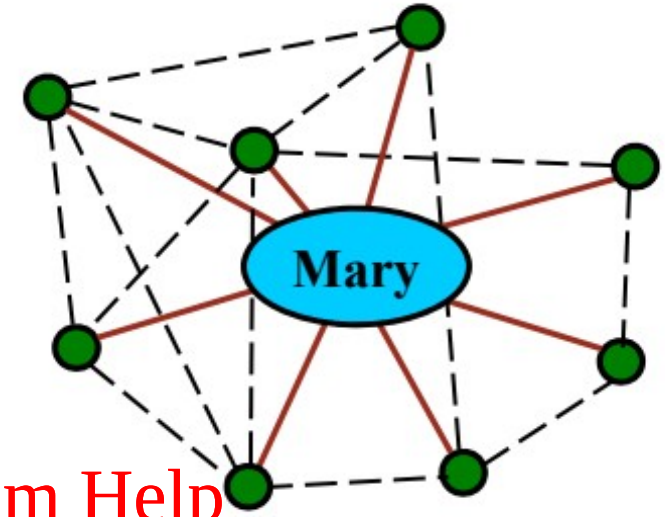
Sampling

- It is often necessary to sample social networks
  - practical reasons
  - computational reasons
- Key point
  - understand what sampling is doing to your data

Egocentric Network Analysis

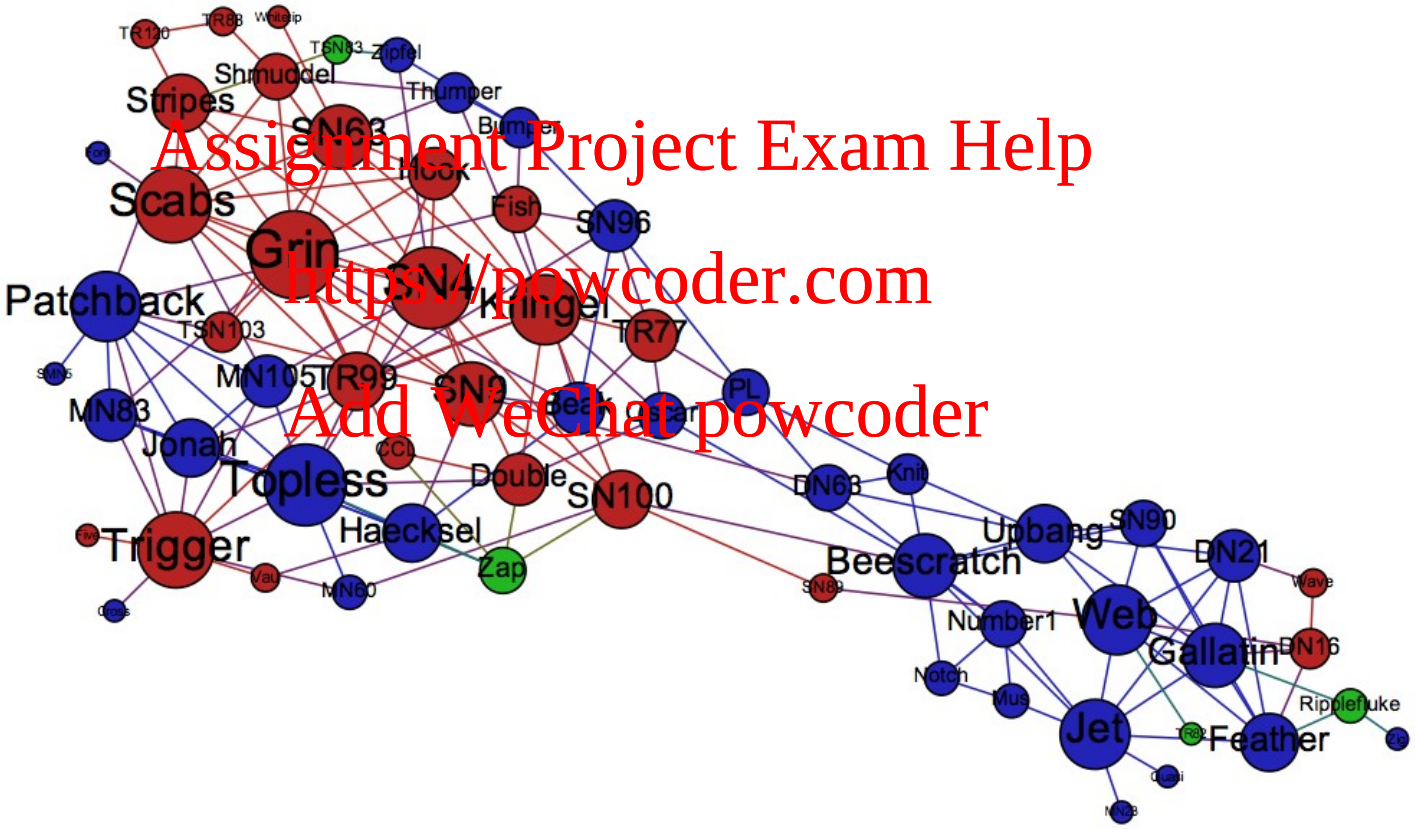
- Instead of looking at the whole network,
  - look at the local networks of some nodes
- A different type of analysis than overall network analysis, but it shows the role of an individual in context.

Ego networks



- Also, First-person network
  - also, 1 neighborhood
  - first-order zone
  - 1.5 neighborhood
- Pieces
  - ego
  - alters
  - ties between alters
- If we don't have ties between alters
  - it isn't a network, just a list of contacts

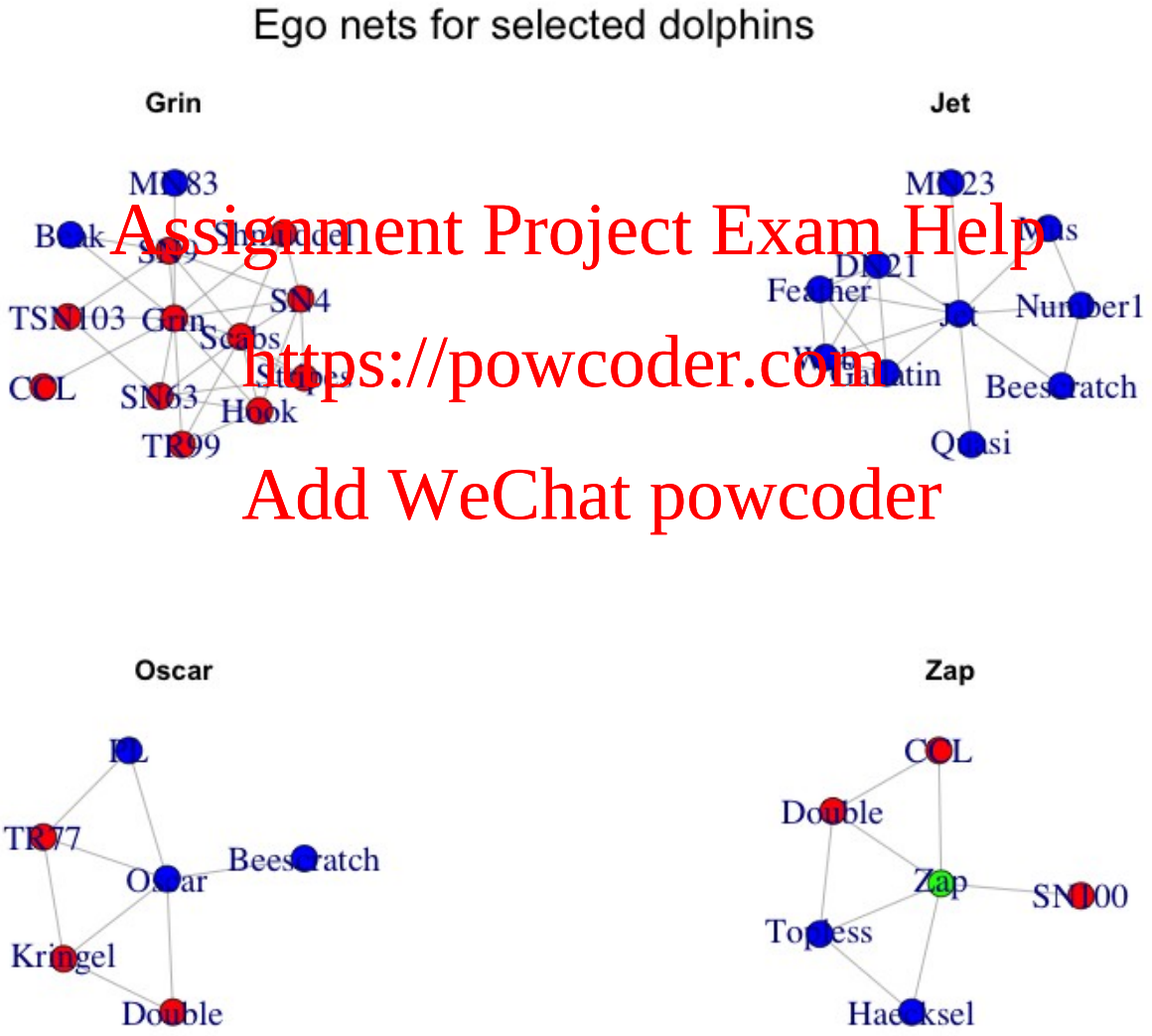
Dolphin network



make_ego_graph

- can specify any distance away from the ego node
- order = 1
  - is usually what is meant

Ego networks



An alternative / complement

- Instead of studying the whole network
  - look at particular individuals
- Useful
  - if network is too large
  - if network is hard to discover
  - if we want to focus on particular individuals

Break