Measurement and methodology for mining mobile, cloud-based, social systems

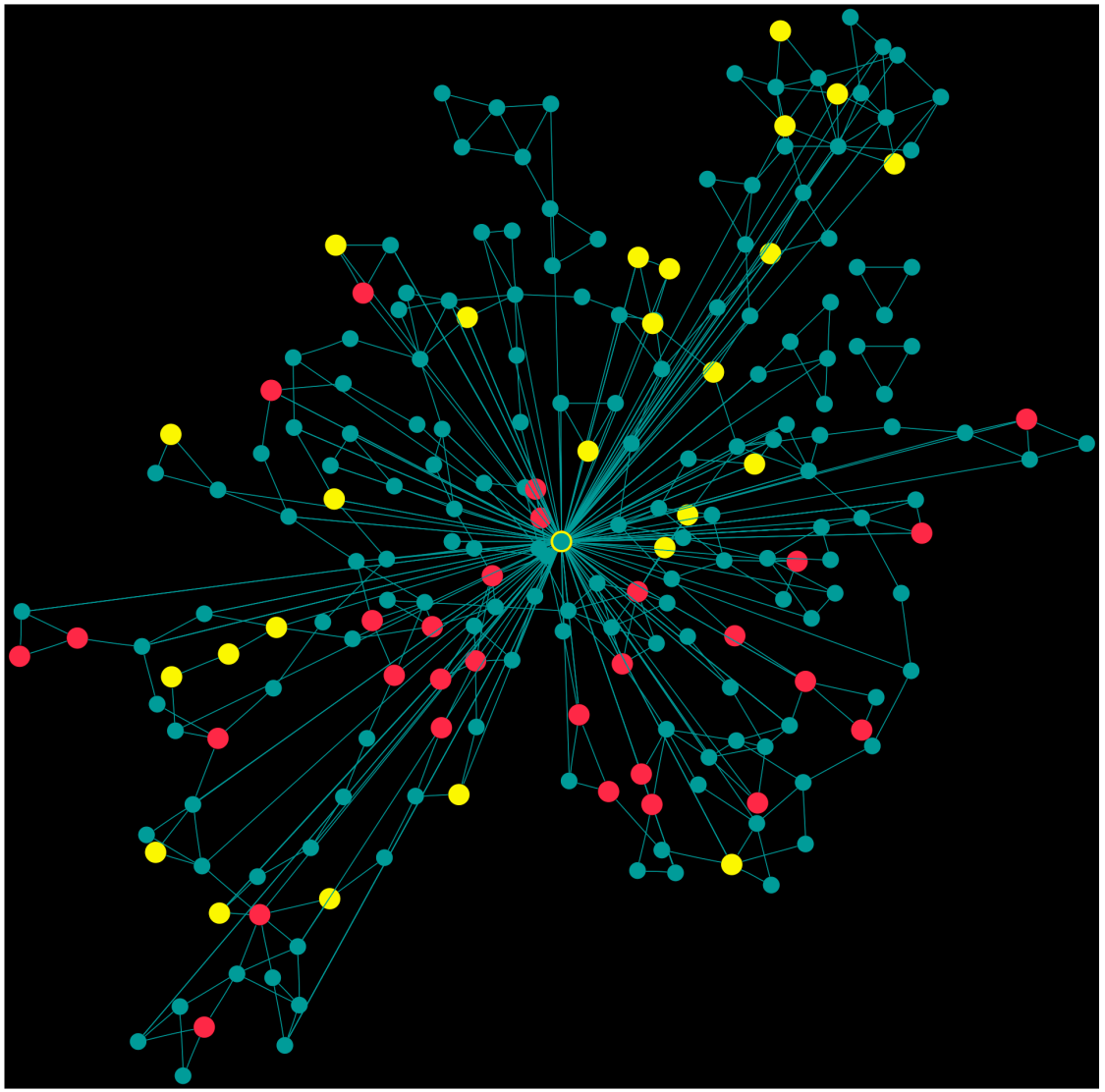
Jennifer Neville
Departments of Computer Science and Statistics
Purdue University

Social network prediction and mining

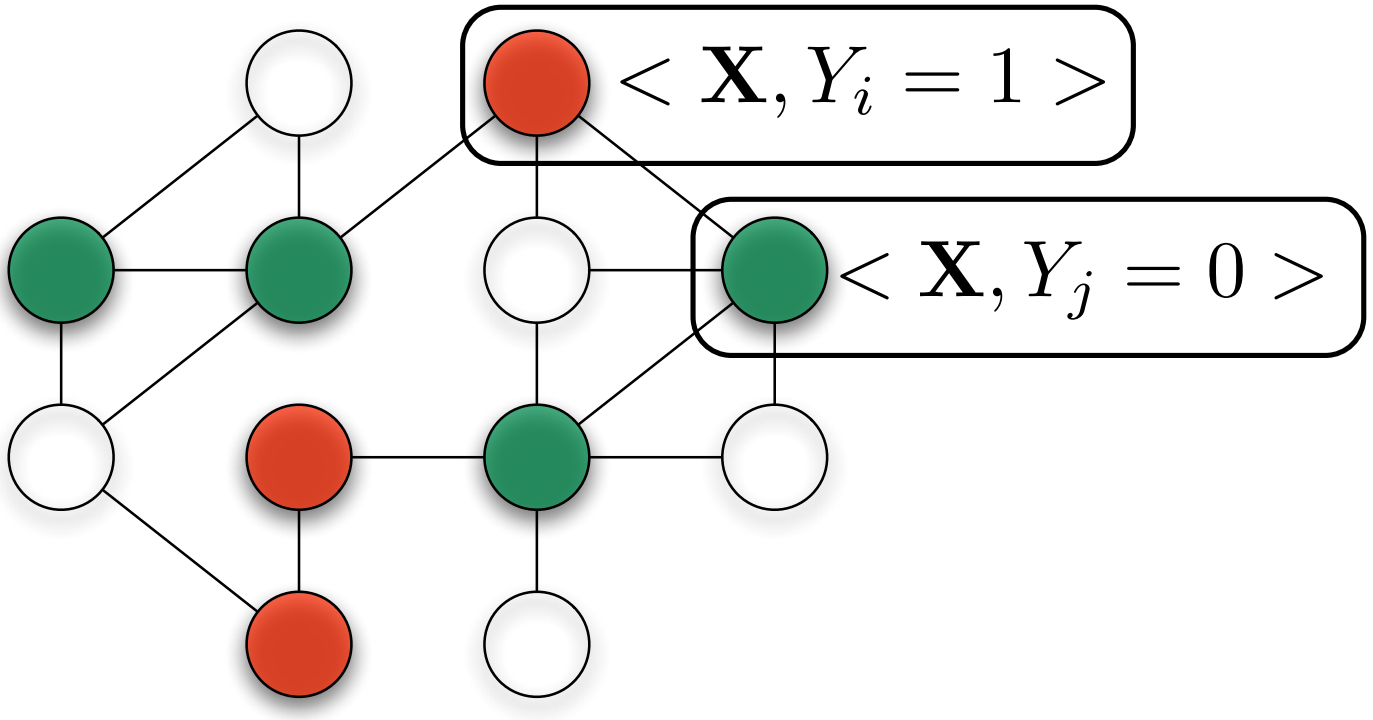
Nodes:
- Facebook users

Edges:
- Articulated friendships

Example task:
- Predict user preferences based on friendships (e.g., political views)



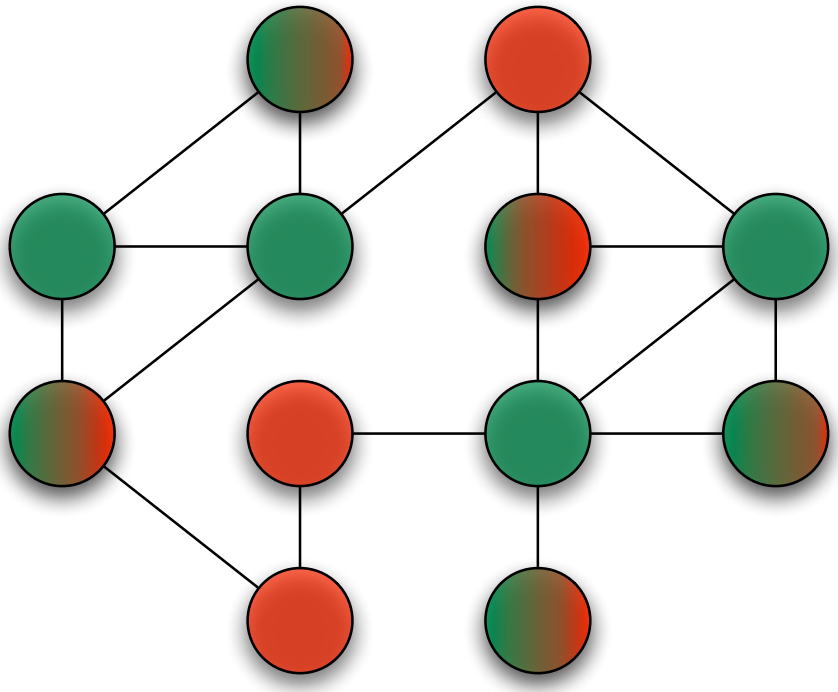
Network prediction models



***Goal**: Estimate joint distribution* $P(\mathbf{Y}|\{\mathbf{X}\}_n, G)$

... or conditional distribution $P(Y_i|\mathbf{X}_i, \mathbf{X}_R, \mathbf{Y}_R)$

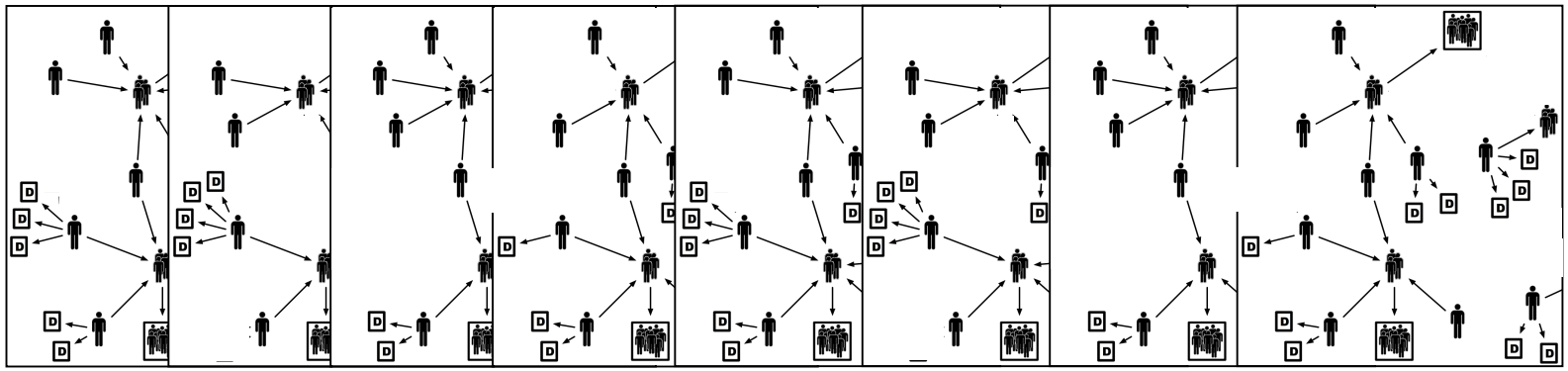
Network prediction models



***Within-network prediction**: Use joint distribution to collectively infer unobserved class labels*

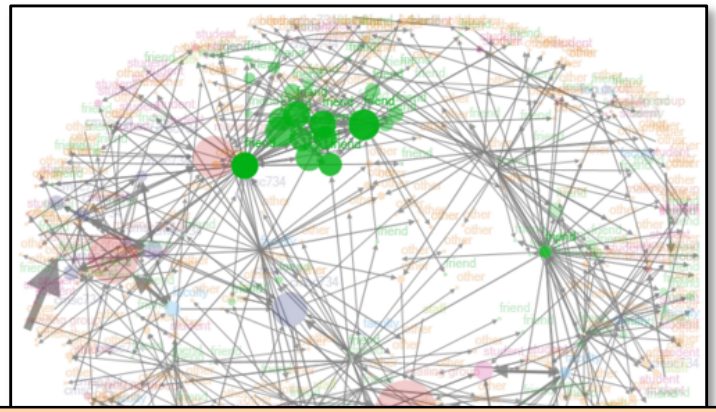
Implicit assumption of many statistical methods

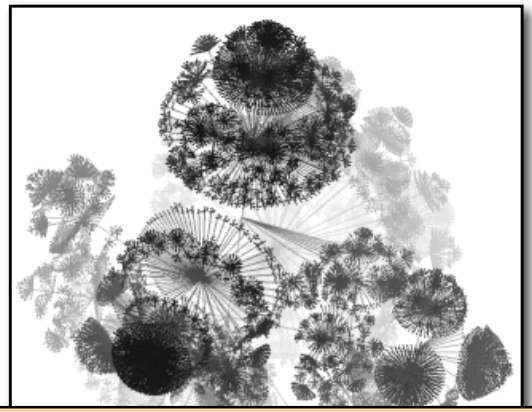
- Domain consists of a population of independent graph samples

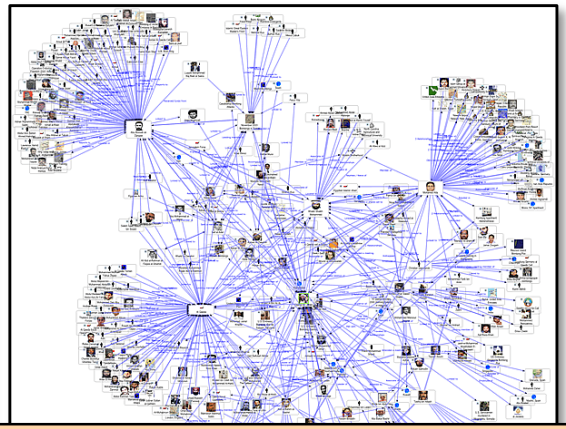


- Increase in data corresponds to acquiring a larger number of graphs
    - When the graphs are independent, we can reason about the characteristics of learning algorithms in the limit... as the number of available graph samples increases

Statistical learning algorithms are often applied to a **single network**

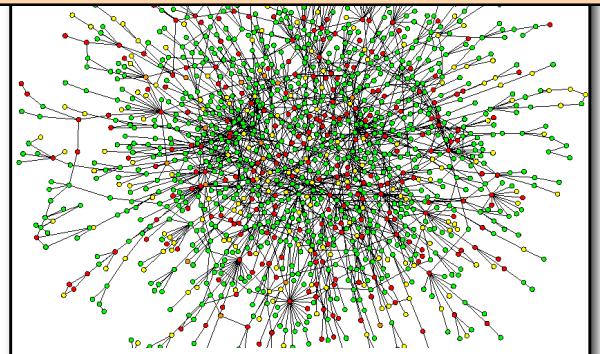




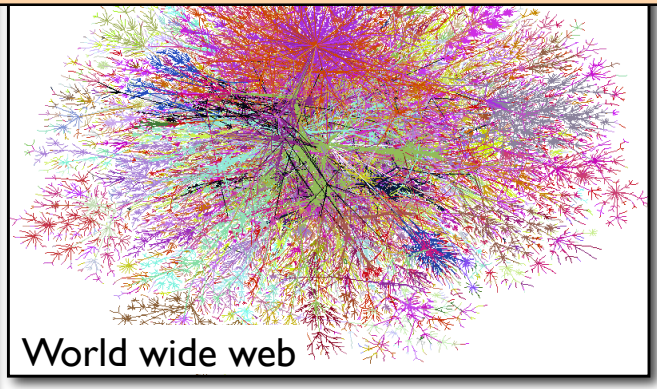


In this case, an increase in dataset size corresponds to acquiring a larger sample from the network...

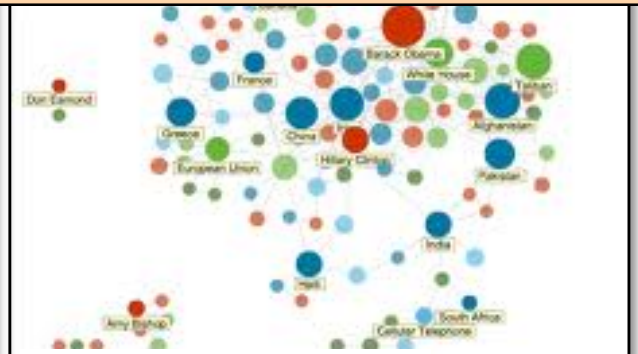
this changes the statistical foundation for analysis and learning



Gene/protein networks



World wide web

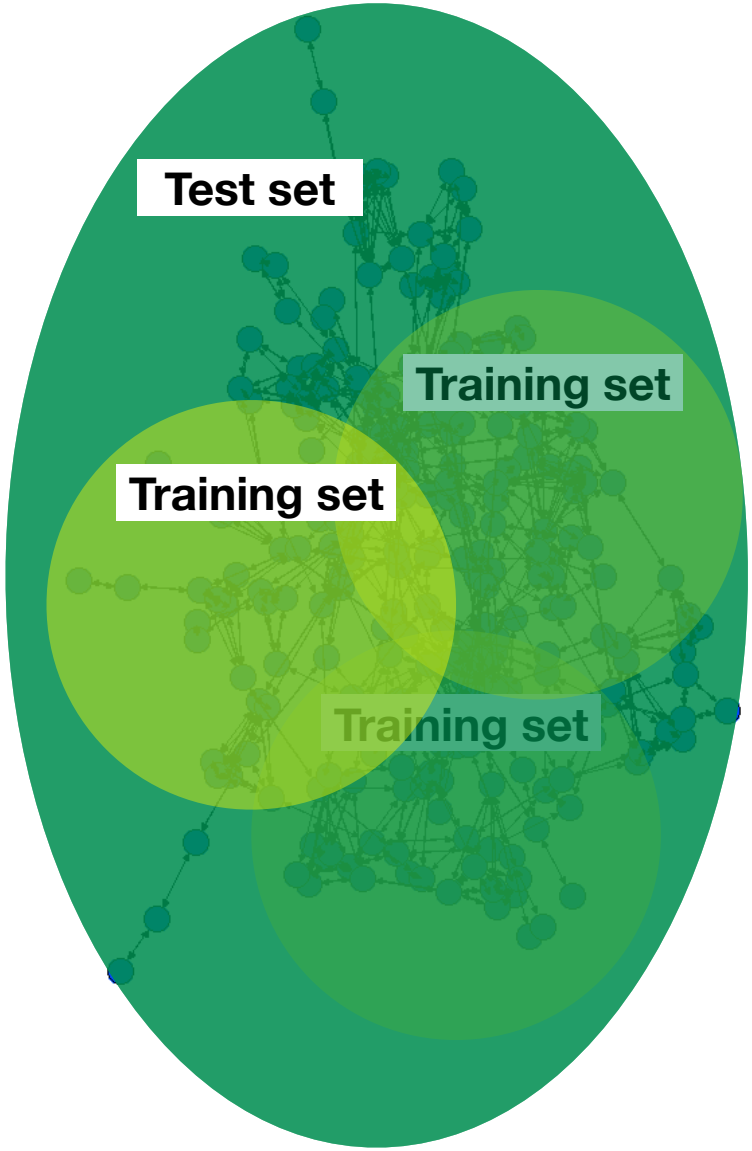


Organizational networks

Observations

- Relational dependencies can significantly improve predictions through the use of collective inference models...
- ...but current methods make assumptions about data and model characteristics that are often not appropriate for many real-world domains
  - Data comprises a single network, *not a population of networks*
  - Relationship information is heterogeneous, *not uniform/stationary*
  - Label and attribute information is sparse, *not fully labeled*
  - Data is dynamically evolving, *not static with respect to time*
- *Need to consider graph/data structure carefully and understand its impact on modeling in order to best exploit the relational information for prediction*
  - In particular, we need to characterize effects analytically... but heterogeneous, dependent, dynamic structure makes this difficult

Example: How do we sample networks when comparing algorithm performance?

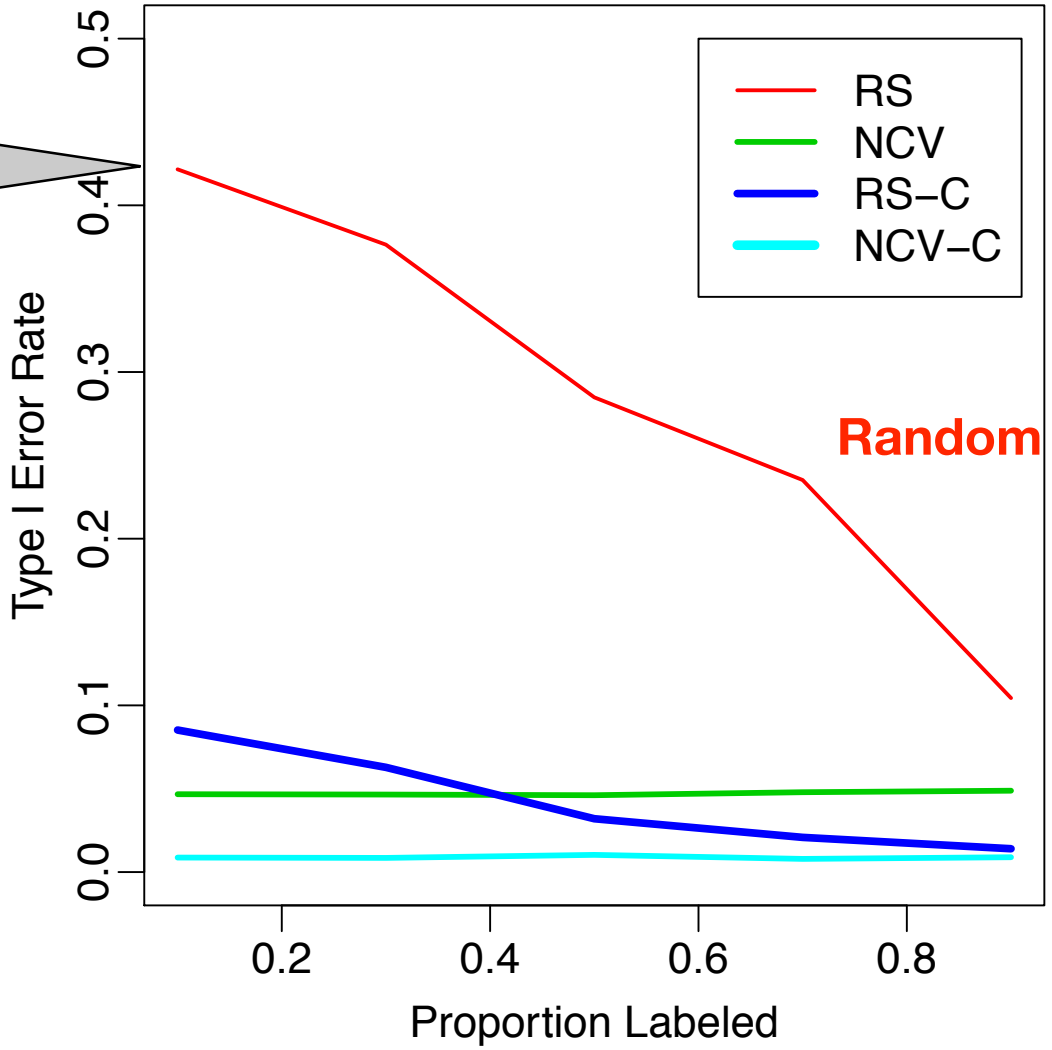


Common approach

Use **repeated random sampling** to create multiple sets of labeled/ unlabeled nodes

Naive sampling results in dependent samples... which leads to high Type I error *(ICDM'09)*

Up to 40% of the time algorithms will appear to be different when they are not!



Random resampling

Random resampling with analytical correction *(ECML'11)*

Challenge 1

- Analyzing data from social networks, social media, and communication can lead to better understanding of behavior in social systems
  ⇒ this can guide the development of more robust and efficient systems

- Data about behavior can indicate how to:
  - Distribute data
  - Predict load
  - Manage data privacy
- Because use will depend on:
  - Social roles
  - Communities
  - Interactions

Challenge 2

- The structure/availability of data make it difficult to accurately test hypotheses about data but the system components facilitate data collection/management ⇒ thus systems can support the development of more accurate statistical tools

- Data characteristics:
  - BIG
  - Long-range dependencies
  - Heterogeneous
  - Dynamic/streaming (“FAST” data)
  - Partially-observable (due to both privacy settings and proprietary access)
  - Behavior may depend on system choices (e.g., ranking results)

- These characteristics are hard to control independently and complicate the methodology for evaluating algorithms and testing hypotheses about data

Open questions

- How to sample?
  - Small samples cannot capture all the properties of large graphs, but is that important to test hypothesis about behavior and algorithms?
  - We have not done a good job of formulating the problem of sampling from these large, dynamic network with events and content.
- When are friendship links more informative than similarity links with unknown users? (e.g., social influence vs. recommendation)
  - Relationships are a noisy indicator or latent (hidden) similarity, if the network is locally sparse is is worth considering?
- How to specify and model the population?
  - Current graph models do not capture the natural variation we see in real datasets (e.g., nodes come/go, connectivity varies)
  - This limits our ability to do anomaly detection and hypothesis testing

Open questions

- Data
  - How to combine data from multiple sources?
  - How to provide shared access to data? Who owns the data?
    Explore secure multi-party computation models that can analyze multiple network data sources without sharing?
  - Access control policies? *Impression management*
- Mobile
  - Usage patterns changes wrt content access, context, and expectation
  - Does this change interactions/behavior among users?
- Cloud
  - Layered, coevolving networks (internet, web, social network)
  - What is the effect of networks overlaid on networks (e.g., social network is filter for news access)?

Research direction

- Data mining/machine learning researchers need to collaborate with systems researchers as well as **social scientists** to understand how user behavior should impact the next generation of systems
  - Analyzing data from social networks, social media, and communication can lead to better understanding of behavior in social systems
    ⇒ this can guide the development of more robust and efficient systems
  - The structure/availability of data make it difficult to accurately test hypotheses about data but the system components facilitate data collection/management
    ⇒ thus systems can support the development of more accurate statistical tools
- **Broader Impacts**:
  - Data collection efforts that can drive computational social science to understand social behavior and dynamics
  - Systems to enhance learning, collaboration, health, etc.