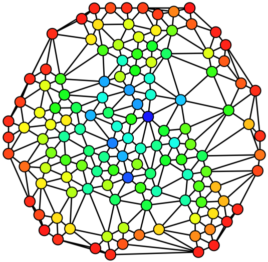
The future of complex networks:

statistics, algorithms and causality

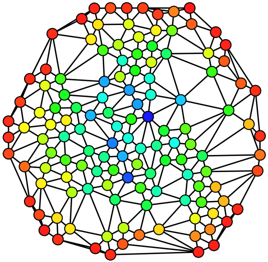
Alexander Volfovsky
Department of Statistical Science, Duke University

October 11, 2017
National Academies: Leveraging Advances in Social Network Thinking for National Security



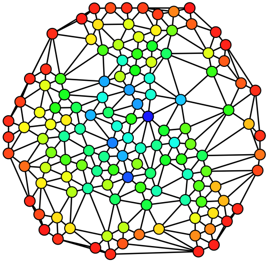
- Networks are everywhere

Image source: Wikimedia



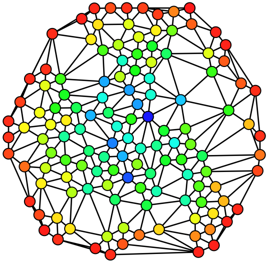
- Networks are everywhere
- Problems of interest:

Image source: Wikimedia



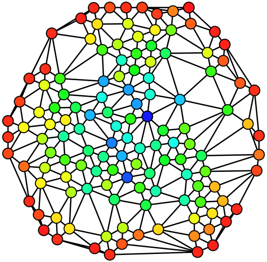
- Networks are everywhere
- Problems of interest:
  - Explaining current ties

Image source: Wikimedia



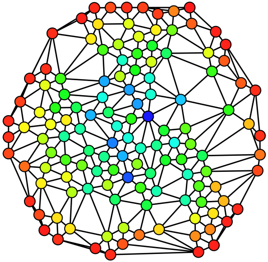
- Networks are everywhere
- Problems of interest:
  - Explaining current ties
  - Predicting future ties

Image source: Wikimedia



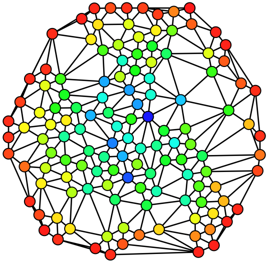
- Networks are everywhere
- Problems of interest:
  - Explaining current ties
  - Predicting future ties
  - Detecting and understanding communities

Image source: Wikimedia



- Networks are everywhere
- Problems of interest:
  - Explaining current ties
  - Predicting future ties
  - Detecting and understanding communities
  - Running experiments on networks

Image source: Wikimedia



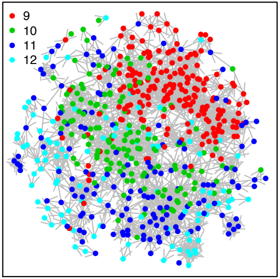
- Networks are everywhere
- Problems of interest:
  - Explaining current ties
  - Predicting future ties
  - Detecting and understanding communities
  - Running experiments on networks

Address **statistical**, **engineering** and **substantive** problems

Image source: Wikimedia

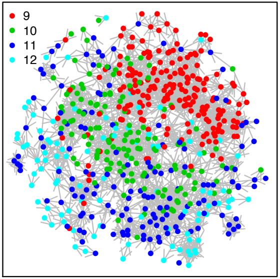
Statistical and substantive

- Datasets: PROSPER, NSCR, AddHealth



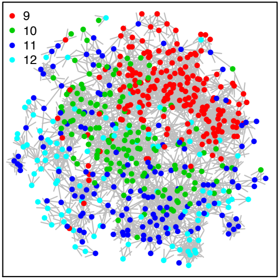
Statistical and substantive

- Datasets: PROSPER, NSCR, AddHealth
- Relate network characteristics to individual-level behavior



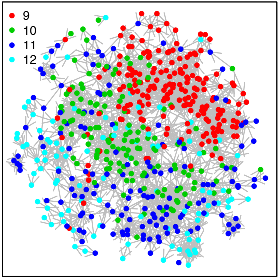
Statistical and substantive

- Datasets: PROSPER, NSCR, AddHealth
- Relate network characteristics to individual-level behavior
- Literature: ERGM, latent variable models



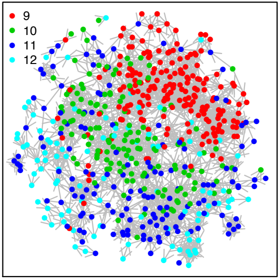
Statistical and substantive

- Datasets: PROSPER, NSCR, AddHealth
- Relate network characteristics to individual-level behavior
- Literature: ERGM, latent variable models
- Assumptions:
  - Data is fully observed
  - The support is the set of all sociomatrices



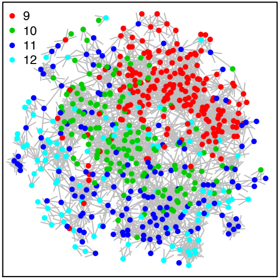
Statistical and substantive

- Datasets: PROSPER, NSCR, AddHealth
- Relate network characteristics to individual-level behavior
- Literature: ERGM, latent variable models
- Assumptions:
  - Data is fully observed
  - The support is the set of all sociomatrices
- In practice:
  - Ranked data
  - Censored observations



Statistical and substantive

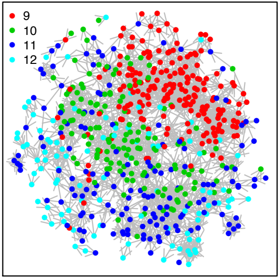
- Datasets: PROSPER, NSCR, AddHealth
- Relate network characteristics to individual-level behavior
- Literature: ERGM, latent variable models
- Assumptions:
  - Data is fully observed
  - The support is the set of all sociomatrices
- In practice:
  - Ranked data
  - Censored observations



Hoff, Fosdick, Volfovsky and Stovel (2013) introduces a likelihood that accommodates the ranked and censored nature of data from Fixed Rank Nomination (FRN) surveys and allows for estimation of regression effects.

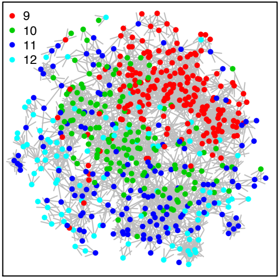
Difficulties that come up

- Communities are frequently based on more than one attribute.



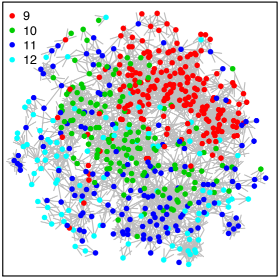
Difficulties that come up

- Communities are frequently based on more than one attribute.
- We can include that in complicated models that require expensive algorithms.



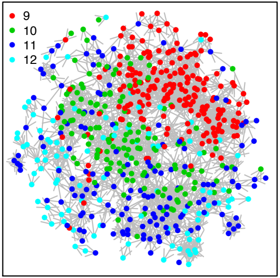
Difficulties that come up

- Communities are frequently based on more than one attribute.
- We can include that in complicated models that require expensive algorithms.
- We can run fast algorithms based on simpler models.



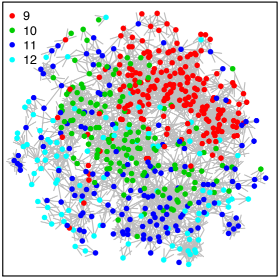
Difficulties that come up

- Communities are frequently based on more than one attribute.
- We can include that in complicated models that require expensive algorithms.
- We can run fast algorithms based on simpler models.
- What happens to fast algorithms under mild misspecification?



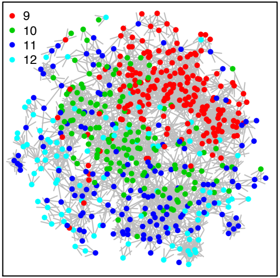
Difficulties that come up

- Communities are frequently based on more than one attribute.
- We can include that in complicated models that require expensive algorithms.
- We can run fast algorithms based on simpler models.
- What happens to fast algorithms under mild misspecification?
- AddHealth friendships might be a stochastic blockmodel plus a bit of noise.



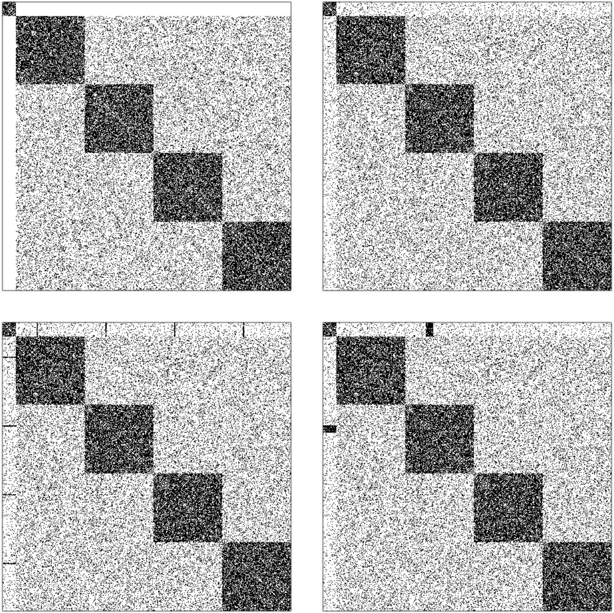
Difficulties that come up

- Communities are frequently based on more than one attribute.
- We can include that in complicated models that require expensive algorithms.
- We can run fast algorithms based on simpler models.
- What happens to fast algorithms under mild misspecification?
- AddHealth friendships might be a stochastic blockmodel plus a bit of noise.

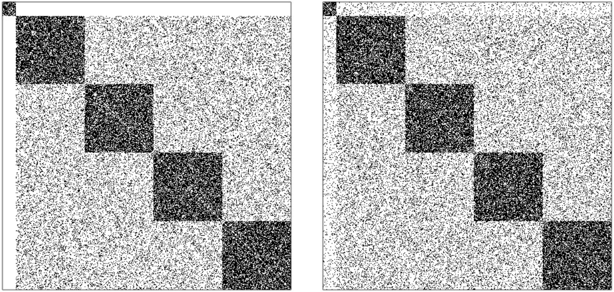


Need new tools to understand

Specific problems: detection

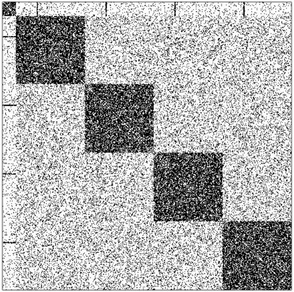


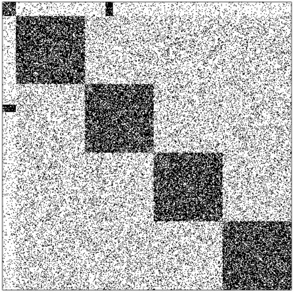
Easy



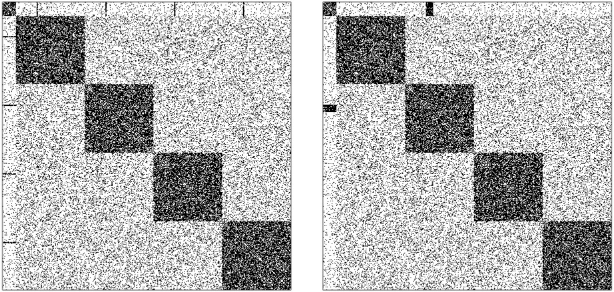
We have fast machinery to do this well
(Spectral methods and guarantees for the stochastic blockmodel)

Hard



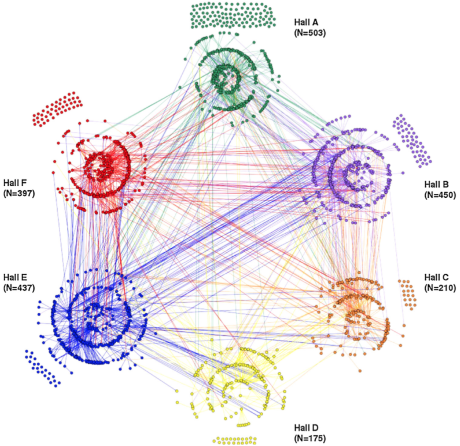


Hard



Looks like multiple or overlapping memberships
We need to build fast machinery to do this

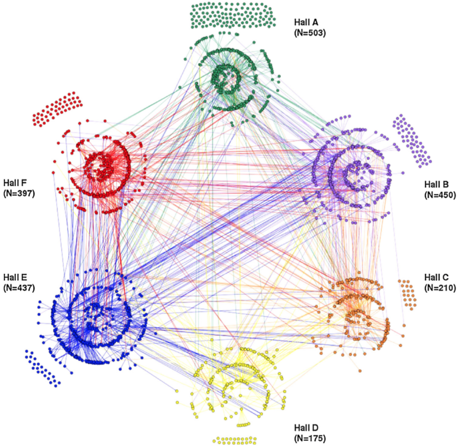
Specific problems: disease spread



- Want to study efficacy of isolation as treatment for influenza-like illness.

Image source: Figure 9 of "Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial" by Aiello et al.

Specific problems: disease spread



- Want to study efficacy of isolation as treatment for influenza-like illness.
- Interested in spread, duration of illness, etc.

Image source: Figure 9 of "Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial" by Aiello et al.

Experimental design with networks

- Want to estimate "causal effects".

Experimental design with networks

- Want to estimate “causal effects”.
- When running experiments, quantity of interest should guide the randomization strategy.

Experimental design with networks

- Want to estimate "causal effects".
- When running experiments, quantity of interest should guide the randomization strategy.
- Total network effect is studied by Eckles, Karrer and Ugander (2014) – they propose graph-cluster randomization.

Experimental design with networks

- Want to estimate "causal effects".
- When running experiments, quantity of interest should guide the randomization strategy.
- Total network effect is studied by Eckles, Karrer and Ugander (2014) – they propose graph-cluster randomization.
- Basse and Airoldi (2017) describe optimal design for the treatment effect under homophily.

Experimental design with networks

- Want to estimate “causal effects”.
- When running experiments, quantity of interest should guide the randomization strategy.
- Total network effect is studied by Eckles, Karrer and Ugander (2014) – they propose graph-cluster randomization.
- Basse and Airoldi (2017) describe optimal design for the treatment effect under homophily.
- Jagadeesan, Pillai and Volfovsky (2017) provide a new graph-based randomization technique for estimating direct effects with arbitrary interference and homophily.

How do we put everything together?

Problems that should be addressed together

- Substantive network based goals:
  - Find someone
  - Learn something about a group
  - Get people (or computers) to do something
- Observed networks are full of uncertainty (statistical problem)
- Available models are too computationally expensive (engineering problem)

Thank you!

Website: `https://volfovsky.github.io/`

- Hoff, Fosdick, Volfovsky and Stovel. Likelihoods for fixed rank nomination networks (2013). Network Science 1 (03), 253-277.
- Jagadeesan, Pillai and Volfovsky. Designs for estimating the treatment effect in networks with interference (2017). arXiv:1705.08524.