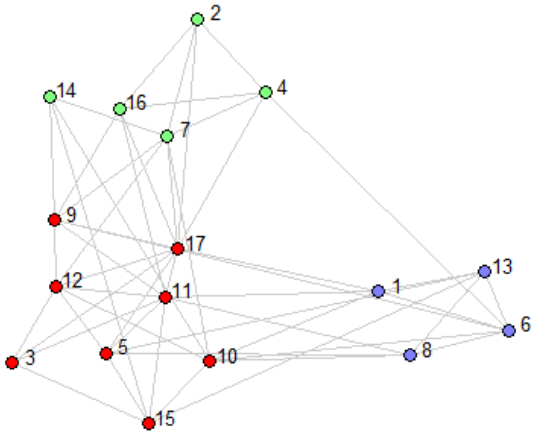
Statistical models for analyzing dynamic social network data

Kevin S. Xu (University of Toledo)

2018/09/28

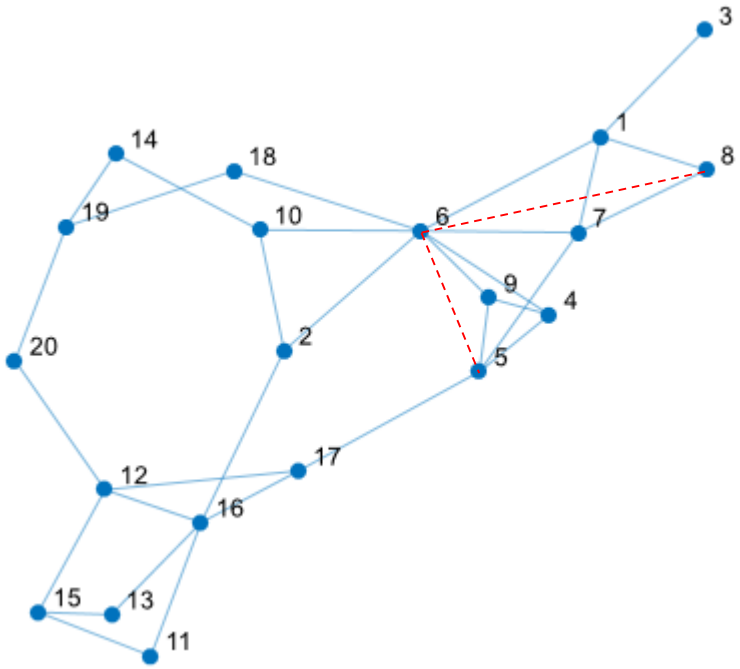


COLLEGE OF ENGINEERING
THE UNIVERSITY OF TOLEDO

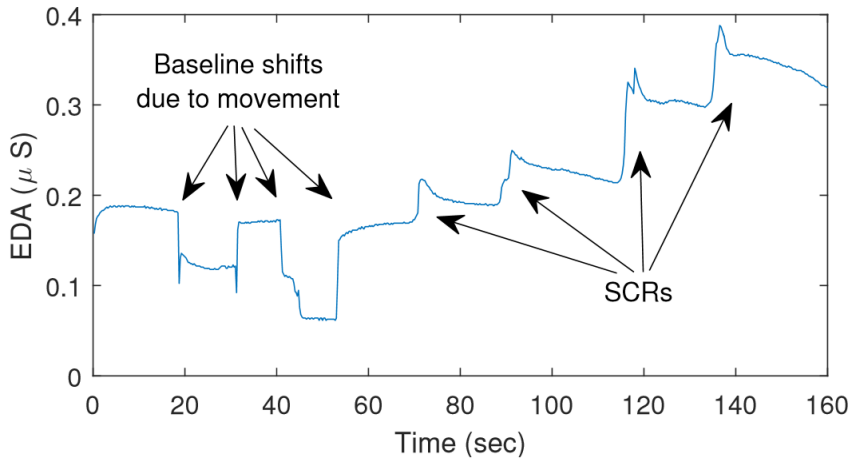
IDEAS Lab research areas

- Interdisciplinary Data Engineering and Science (IDEAS) Lab at University of Toledo

Network science



Wearable data analytics



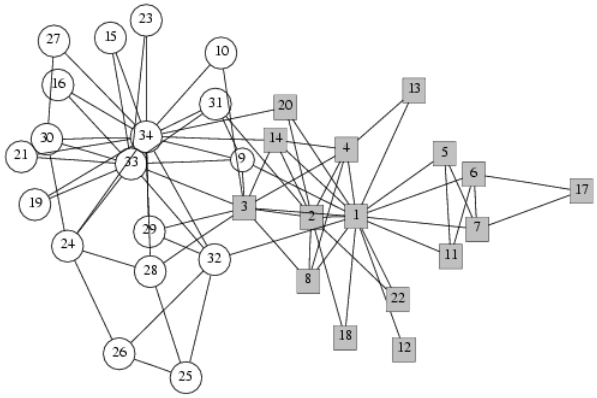
Computational human dynamics

- Advances in technology enable the study of human dynamics on a much larger scale with finer resolution
  - Portable sensors capable of continuously collecting data on an individual's movement, activities, interactions, and responses to external stimuli in free-living settings
- What can we learn and predict about dynamics of human behavior from these data?
  - Need to develop new statistical models and algorithms to analyze new modalities of data

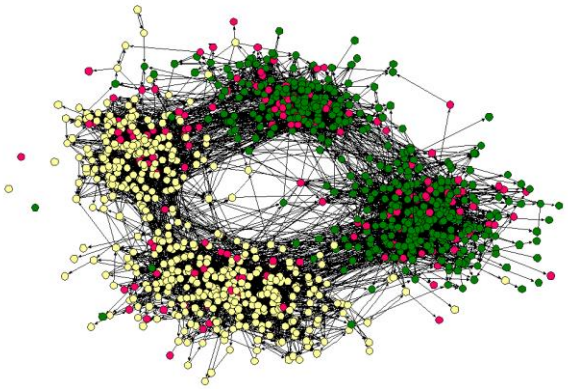
THE UNIVERSITY OF TOLEDO 1872

Social networks

- Wealth of information on human interactions are embedded in social network data
  - Diffusion of information and spreading of diseases
  - Roles, influences, preferences, activity patterns
- Significant interdisciplinary effort dedicated to analyzing social network data
  - Development of statistical models for social network data is an important step to understanding human behavior at the network level



Zachary's Karate club
(Zachary, 1977)



School friendships
(Moody, 2001)

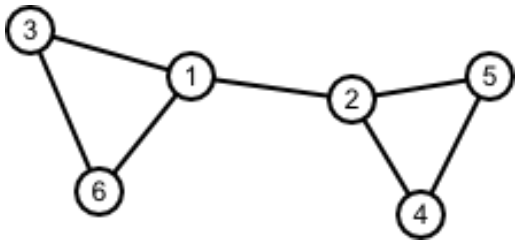
THE UNIVERSITY OF TOLEDO 1872

Outline

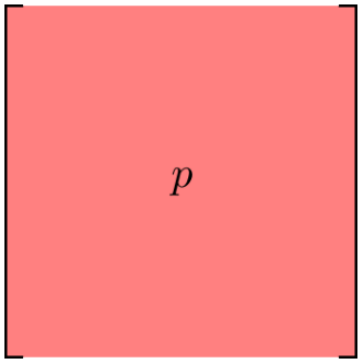
- Statistical models for static networks
  - Stochastic block model (SBM)
- Models for discrete-time dynamic networks
  - Hidden Markov SBM
  - Stochastic block transition model
- Models for continuous-time dynamic networks
  - Block point process model

Statistical models for networks

- Represent network by $n \times n$ adjacency matrix W
  - $w_{ij} = 1$ if there is an edge from node i to j
  - $w_{ij} = 0$ otherwise
  - $w_{ij} = w_{ji}$ for undirected network



$$W = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$



- Erdős-Rényi or $G(n, p)$ model: all edges between nodes $i \neq j$ formed independently with probability p
  - Nodes are homogeneous
  - Too simple (inflexible) to represent many phenomena

Stochastic block model (SBM)

- Stochastic block model (SBM): commonly used model for static networks formalized by Holland et al. (1983)

Definition: A random adjacency matrix W is generated according to an SBM w.r.t. to class vector $\mathbf{c}$ iff

1. For $i \neq j, w_{ij}$ are statistically independent
2. If i and i' are in same class and j and j' are in same class then w_{ij} and $w_{i'j'}$ are identically distributed

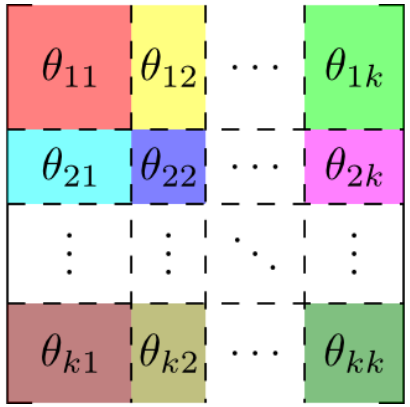
$$W \sim \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1k} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{k1} & \theta_{k2} & \cdots & \theta_{kk} \end{bmatrix}$$

- Membership vector $\mathbf{c}$ (length n)
- Matrix of edge probabilities: Θ $(k \times k)$
- Edges between nodes $i \neq j$ formed independently with probability $\theta_{c_i c_j}$

THE UNIVERSITY OF TOLEDO 1872

Estimators for SBM

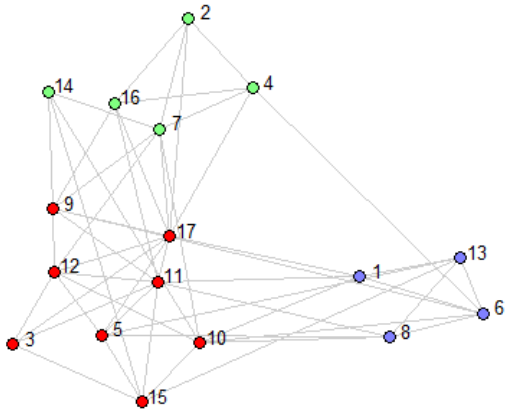
- A posteriori estimation: estimate both $\mathbf{c}$ and Θ from W
- Maximum-likelihood estimator (MLE)
  - NP-hard: class membership vector is discrete
  - Scales to ~ 20 nodes
- Approximate methods:
  - Markov chain Monte Carlo
    - Scales to ~ 100 nodes
  - Variational expectation-maximization (EM)
    - Scales to ~ 1,000 nodes
  - Spectral clustering
    - Scales to ~ 10,000 nodes
  - All shown to be consistent estimators as $n \to \infty$ as long as network isn't too sparse

$W \sim$ 

THE UNIVERSITY OF TOLEDO 1872

From static to dynamic networks

- A single (static) network is insufficient for representing many complex phenomena
  - Unable to capture temporal dynamics
- Solution: analyze dynamic network representations
- How should we model a dynamic network?



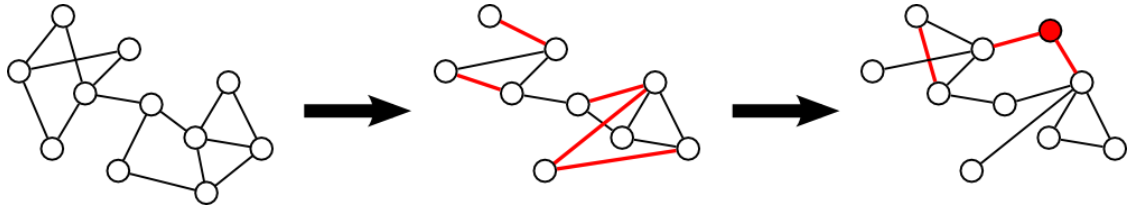
Dynamic social network
(Nordlie, 1958; Newcomb, 1961)

Outline

- Statistical models for static networks
  - Stochastic block model (SBM)
- Models for discrete-time dynamic networks
  - Hidden Markov SBM
  - Stochastic block transition model
- Models for continuous-time dynamic networks
  - Block point process model

Discrete-time dynamic networks

- Network snapshots at discrete time steps
  - Nodes and edges can both appear and disappear over time



  - Adjacency matrix dimensions may change over time

$$W^{t-1} = \begin{bmatrix} 0 & 1 & \dots & 0 \\ 1 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 0 \end{bmatrix} \quad W^{t} = \begin{bmatrix} 0 & 1 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \quad W^{t+1} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 1 \\ 1 & 1 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \dots & 0 \end{bmatrix}$$

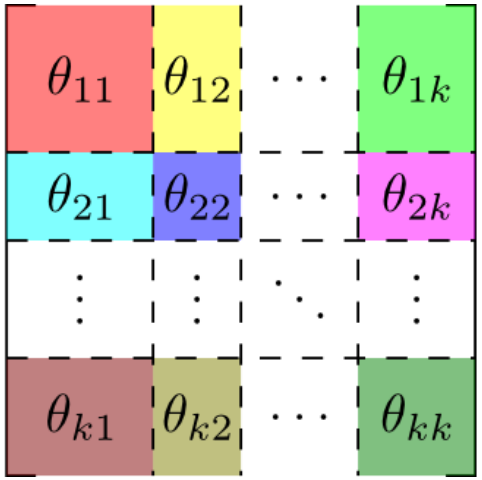
- Also known as network panel data

THE UNIVERSITY OF TOLEDO 1872

Dynamic network models

- Build upon commonly used models for static networks, e.g. stochastic block model (SBM)

Definition: A random sequence of adjacency matrices $W^{1:T}$ is generated according to a dynamic SBM w.r.t. to class vectors $\mathbf{c}^{1:T}$ iff, for each time t, each snapshot W^t is generated according to a static SBM with respect to $\mathbf{c}^t$.

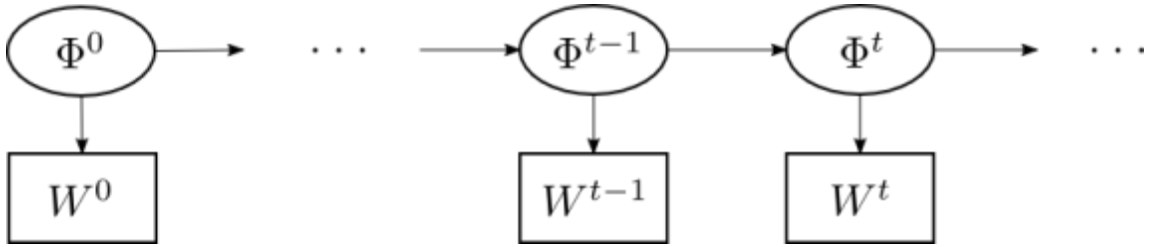
$W \sim$ 

- Incorporate time dependence on model parameters to build more powerful models for (discrete-time) dynamic networks

THE UNIVERSITY OF TOLEDO 1872

Hidden Markov SBM

- First attempt: assume hidden Markov structure
  - Network parameters follow Markov dynamics
  - Each snapshot generated using static network model, e.g. SBM



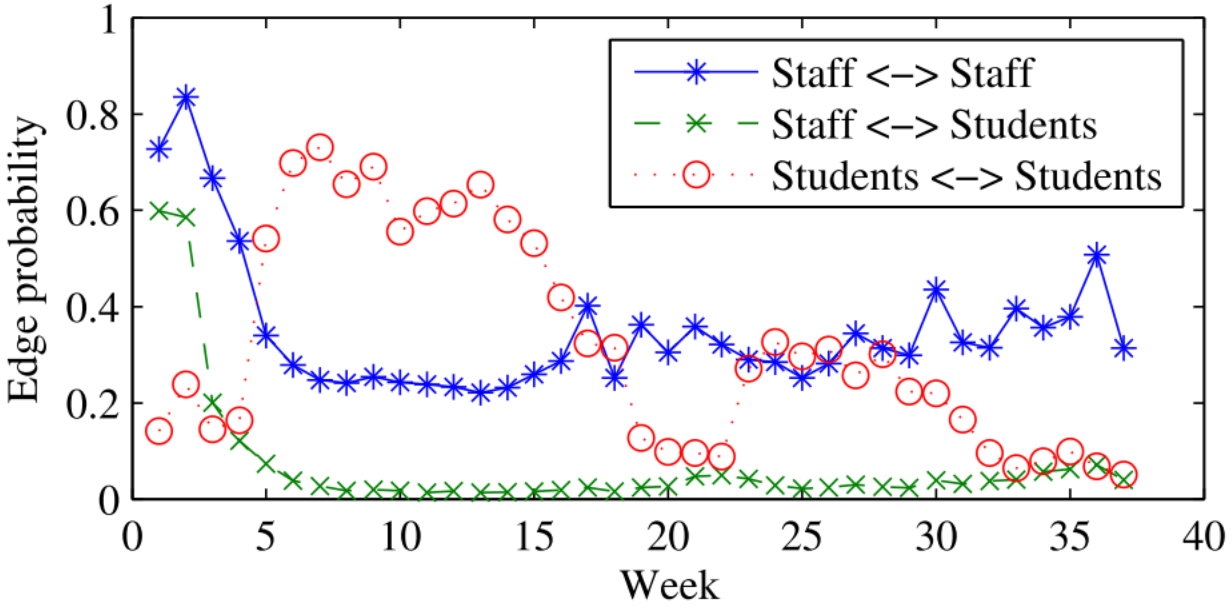
- We propose a hidden Markov SBM (HM-SBM) that allows both class memberships and edge probabilities to vary over time (Xu and Hero III, 2014)
  - Recently extended by Matias and Miele (2017)
- Time dependence in model further complicates inference
  - We exploit an asymptotic Gaussian property for efficient inference using hill climbing + extended Kalman filter
  - Scales to ~ 1,000 nodes

Xu, K. S., & Hero III, A. O. (2014). Dynamic stochastic blockmodels for time-evolving social networks. *IEEE J. Sel. Top. Signal Process.*, *8*(4), 552–562.
Matias, C., & Miele, V. (2017). Statistical clustering of temporal networks through a dynamic stochastic block model. *J. Royal Stat. Soc. Ser. B (Stat. Method.)*, 79, 1119-1141.

THE UNIVERSITY OF TOLEDO 1872

MIT Reality Mining network

- Dynamic social network of physical proximity using Bluetooth between 94 students and staff at MIT over 2 semesters (Eagle et al., 2009)
- HM-SBM reveals differences in temporal dynamics of 2 classes

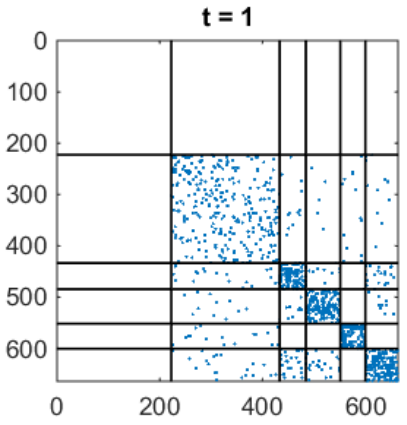


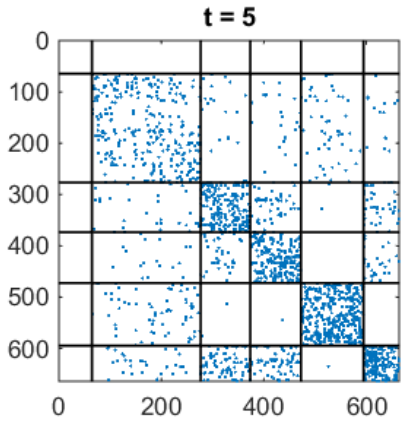
N. Eagle et al. Inferring friendship network structure by using mobile phone data. *PNAS*, 106(36):15274–15278, 2009.

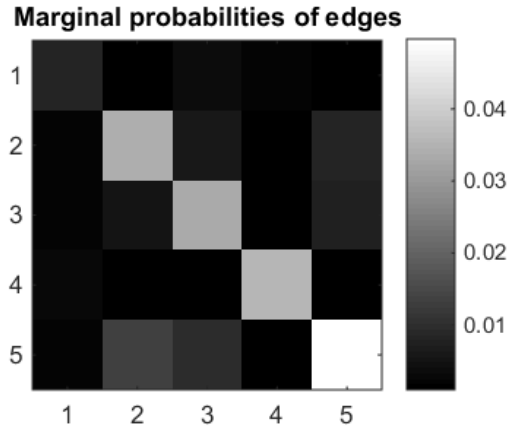
THE UNIVERSITY OF TOLEDO 1872

Facebook wall posts

- Dynamic directed network of Facebook wall posts from > 60,000 users over > 2 years (Viswanath et al., 2009)
  - We use a subset with ~700 nodes, 9 time steps (90 days each), and 5 classes
  - Edge $i \rightarrow j$ at time t denotes that user j posted on Facebook wall of user i at least once in time step t





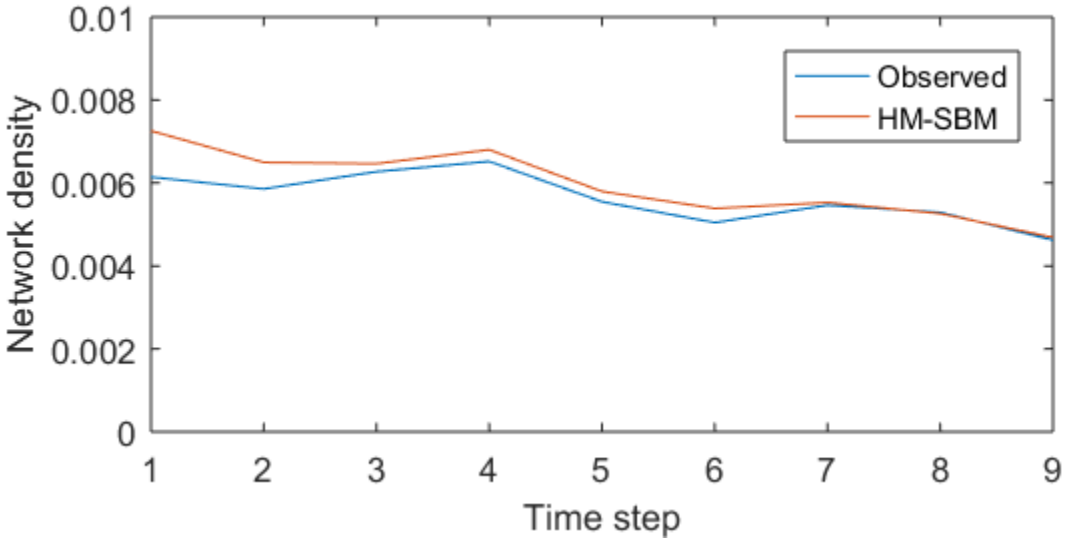


Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. (2009). On the evolution of user interaction in Facebook. In *Proc. 2nd ACM Workshop Online Soc. Netw.*

THE UNIVERSITY OF TOLEDO 1872

Posterior predictive checks

- Do networks generated from the model differ from the observed network in ways we care about?
- Posterior predictive check: Compare test statistic computed on observed network to test statistic computed on simulated networks
- Test statistic: densities of network snapshots

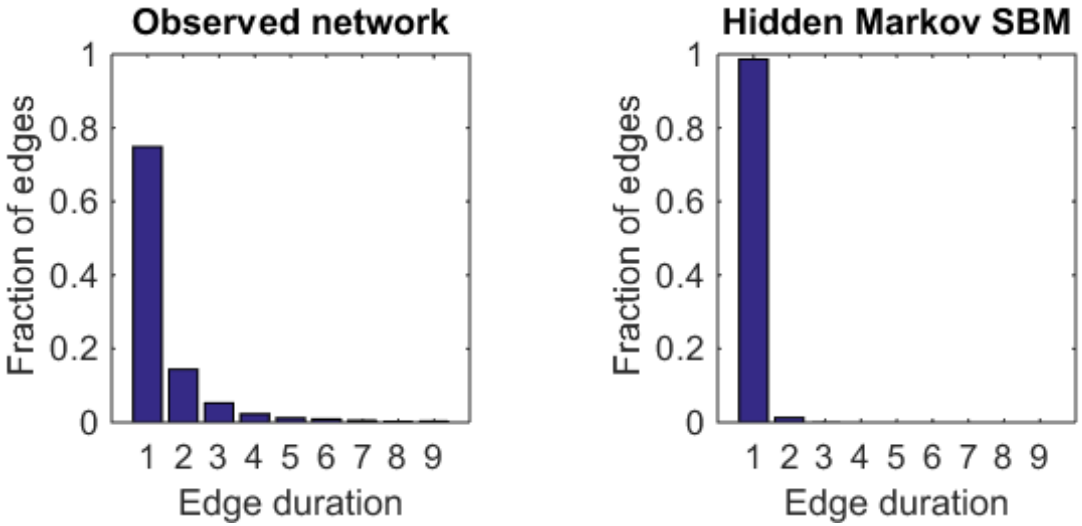


- HM-SBM replicates densities of snapshots reasonably well

THE UNIVERSITY OF TOLEDO 1872

Posterior predictive checks

- Do networks generated from the model differ from the observed network in ways we care about?
- Posterior predictive check: Compare test statistic computed on observed network to test statistic computed on simulated networks
- Test statistic: edge durations

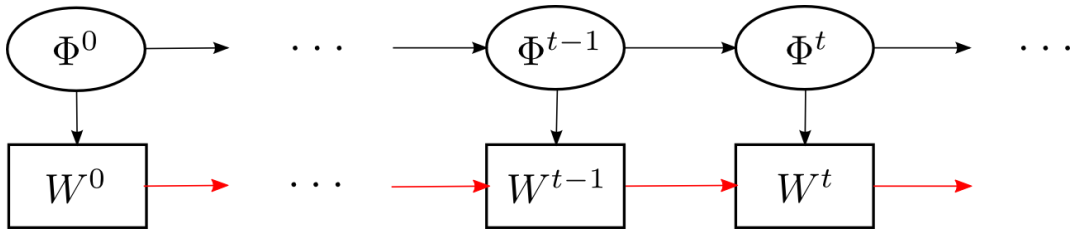


- HM-SBM cannot replicate long-lasting edges in sparse blocks

THE UNIVERSITY OF TOLEDO 1872

Beyond hidden Markov networks

- Hidden Markov structure is tractable but not very realistic assumption in social interaction networks
  - Interaction between two people does not influence future interactions
- Proposed model: Allow current snapshot to depend on current parameters and previous snapshot



  - LFP (Heaukulani and Ghahramani, 2013) and link persistence (Frial et al., 2016) models also satisfies this property
- I propose a stochastic block transition model (SBTM) based on a dynamic SBM
  - Retains asymptotic Gaussian property for efficient inference using hill climbing + EKF
  - Still scales to ~ 1,000 nodes

Heaukulani, C., & Ghahramani, Z. (2013). Dynamic probabilistic models for latent feature propagation in social networks. In *Proc. 30th Int. Conf. Mach. Learn.*

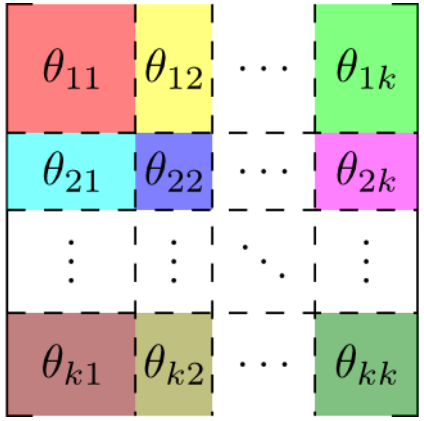
Friel, N., Rastelli, R., Wyse, J., & Raftery, A. E. (2016). Interlocking directorates in Irish companies using a latent space model for bipartite networks. *PNAS*, 113, 201606295.

THE UNIVERSITY OF TOLEDO 1872

Stochastic block transition model

- Main idea: parameterize each block (a, b) with two probabilities
  - Probability of forming new edge $\pi_{ab}^{t|0} = \Pr\left(w_{ij}^t = 1 | w_{ij}^{t-1} = 0\right)$
  - Probability of existing edge re-occurring $\pi_{ab}^{t|1} = \Pr\left(w_{ij}^t = 1 | w_{ij}^{t-1} = 1\right)$

$W \sim$

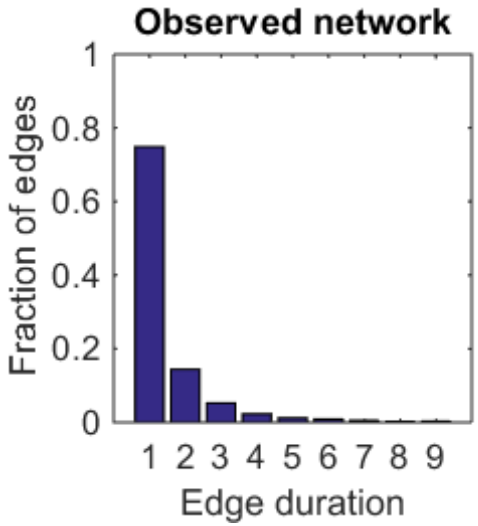


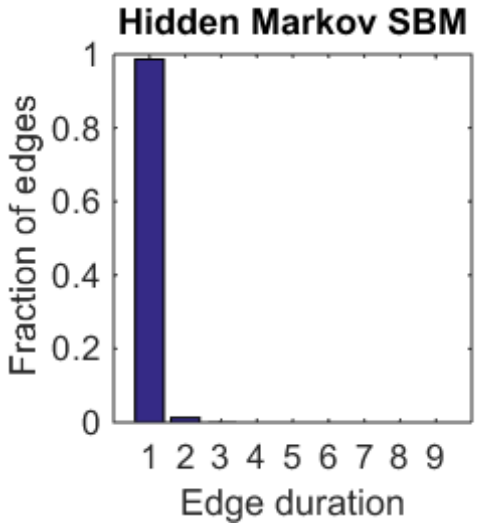
Definition: A random sequence of adjacency matrices $W^{1:T}$ is generated according to an SBTM w.r.t. to class vectors $\mathbf{c}^{1:T}$ iff

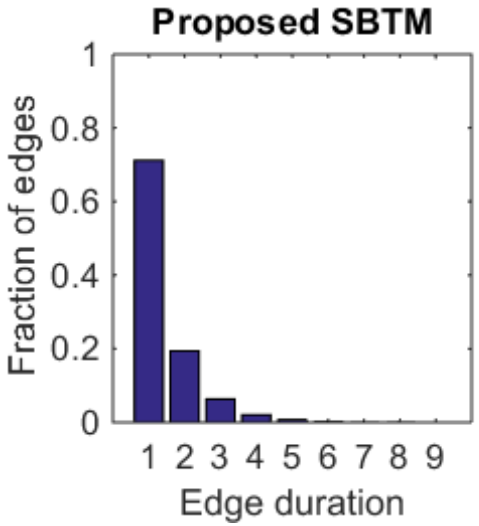
1. The initial adjacency matrix W^1 follows an SBM w.r.t. $\mathbf{c}^1$
2. At any time t, for $i \neq j$, w_{ij}^t are statistically independent
3. At time $t \geq 2$, $\Pr\left(w_{ij}^t = 1 | w_{ij}^{t-1} = 0\right) = \xi_{ij}^t \pi_{ab}^{t|0}$ and $\Pr\left(w_{ij}^t = 1 | w_{ij}^{t-1} = 1\right) = \xi_{ij}^t \pi_{ab}^{t|1}$

Posterior predictive checks

- Do networks generated from the model differ from the observed network in ways we care about?
- Posterior predictive check: Compare test statistic computed on observed network to test statistic computed on simulated networks
- Test statistic: edge durations





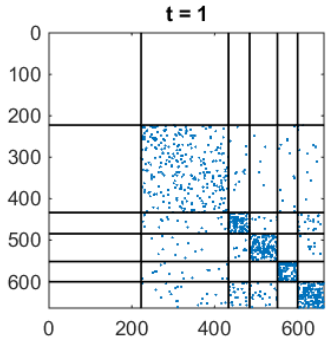


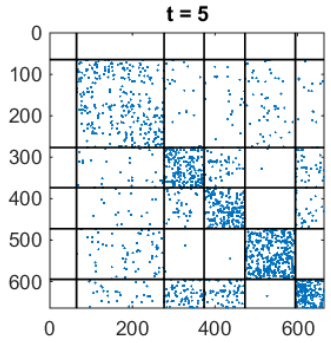
- SBTM can replicate long-lasting edges in sparse blocks

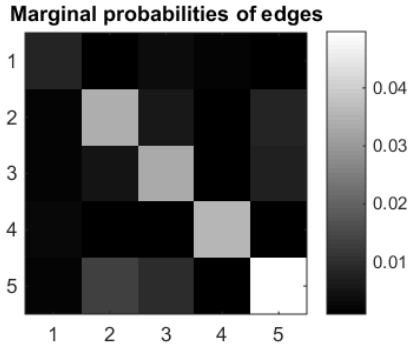
THE UNIVERSITY OF TOLEDO 1872

Behaviors of different classes

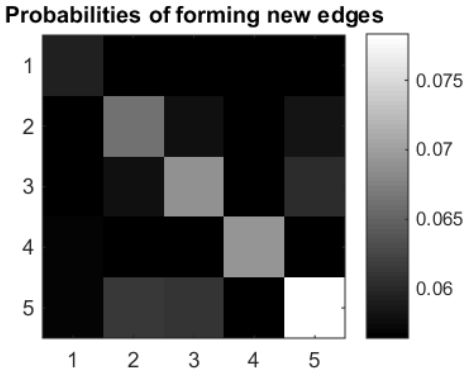
- SBTM retains interpretability of SBM at each time step

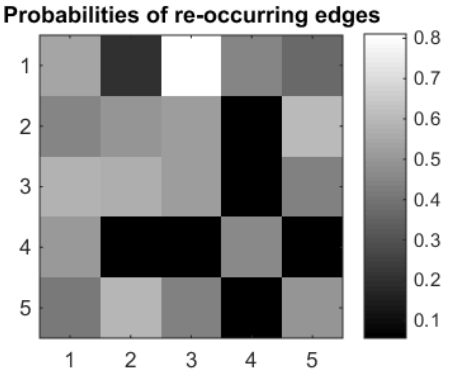






- Q: Do different classes behave differently in how they form edges?





- A: Only for probability of existing edges re-occurring
- New insight revealed by having separate probabilities in SBTM

Xu, K. S. (2015). Stochastic block transition models for dynamic networks. In *Proc. AISTATS*.

THE UNIVERSITY OF TOLEDO 1872

Outline

- Statistical models for static networks
  - Stochastic block model (SBM)
- Models for discrete-time dynamic networks
  - Hidden Markov SBM
  - Stochastic block transition model
- Models for continuous-time dynamic networks
  - Block point process model

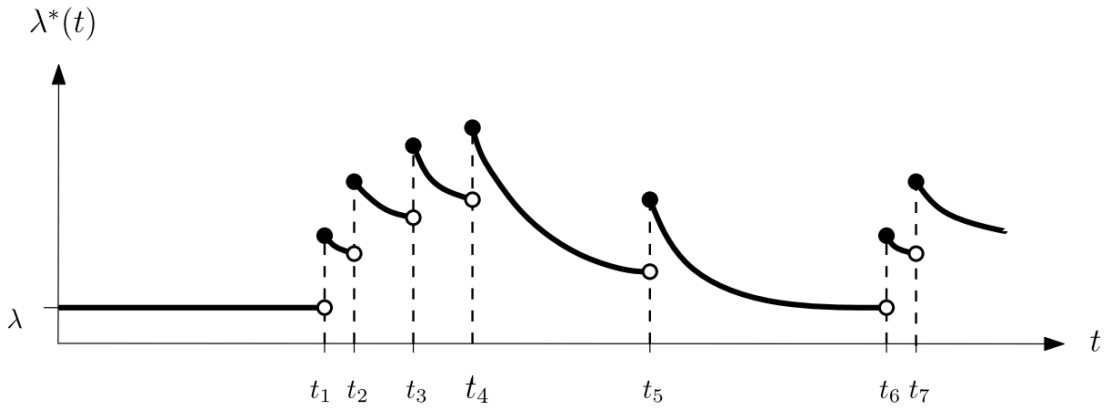
Continuous-time event-based dynamic networks

- Relational event data with fine-grained timestamps
  - Facebook wall posts (Viswanath et al., 2009)
- Represent events as triplets (i, j, t)
- Goal: build statistical model for these relations over time
  - Without aggregating to form discrete-time snapshots

Sender	Receiver	Timestamp
1595	1021	1100626783
4581	5626	1100627183
3806	991	1100640075
521	533	1100714520
521	3368	1100716404
8734	527	1100724840
1017	1015	1100828851
17377	1021	1100832283
2926	726	1100838067

The Block Point Process Model (BPPM)

- Our approach: Model event triplets (i, j, t) directly using SBM-like generative structure
  - Divide nodes into K classes forming $p = K^2$ block pairs (assuming directed events)
  - Generate times of events in each block pair using a point process model
  - Randomly associate event with a pair of nodes (i, j) in the block pair (thinning)
  - We use exponential Hawkes processes in practice



THE UNIVERSITY OF TOLEDO 1872

Adjacency matrix representation

- Construct adjacency matrix $A = A^{[t_1,t_2)}$ from event matrix E
  - $a_{ij} = 1$ if at least 1 event from i to j in $[t_1, t_2)$
- If E follows a BPPM, does the adjacency matrix follow an SBM?
  - If not, can we still use inference techniques for SBM to fit BPPM?

Sender	Receiver	Timestamp
1	2	0.1
2	3	0.4
3	2	0.6
1	2	1.2
1	3	1.3
2	1	1.6

$$A^{[0,1)} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$A^{[1,2)} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Relationship to SBM

- Identical distribution of adjacency matrix entries within block satisfied by BPPM generative procedure
- But independence of entries is not satisfied!
  - Denote deviation from independence by

$$\delta_0 = \Pr(a_{ij} = 0 | a_{i'j'} = 0) - \Pr(a_{ij} = 0)$$
$$\delta_1 = \Pr(a_{ij} = 0 | a_{i'j'} = 1) - \Pr(a_{ij} = 0)$$

Theorem (Asymptotic Independence Theorem). *Consider an adjacency matrix A constructed from the BPPM over some time interval $[t_1, t_2)$. Then, for any two entries a_{ij} and $a_{i'j'}$ both in block b, the deviation from independence given by δ_0, δ_1 defined in (1) is bounded in the following manner:*

$$|\delta_0|, |\delta_1| \leq \min\{1, \mu_b / n_b\}$$

$O(N^2)$ for fixed K

where μ_b denotes the expected number of events in block b in $[t_1, t_2)$, and n_b denotes the size of block b. In the limit as the block size $n_b \to \infty$, $\delta_0, \delta_1 \to 0$ provided μ_b is fixed or growing at a slower rate than n_b. Thus a_{ij} and $a_{i'j'}$ are asymptotically independent in the block size n_b.

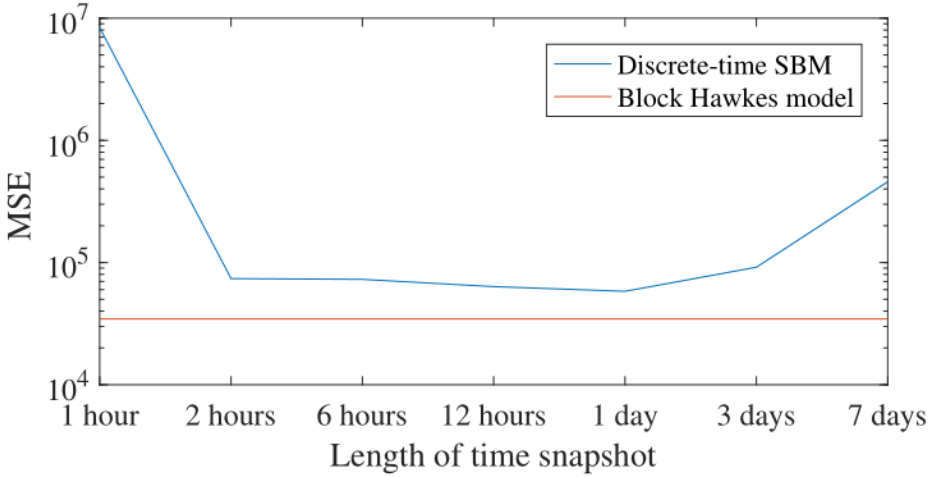
Implications of asymptotic independence

- First result linking point process network models with static network models
- Main implication: for large networks that are not too dense, class estimation methods that work for SBM should also work for BPPM
  - Maximum likelihood estimation and spectral clustering are both consistent as $N \to \infty$ for polylog expected degree (Bickel et al., 2013; Lei and Rinaldo, 2015)
    - Results in $\frac{\mu_b}{n_b} = O\left(\frac{N \operatorname{poly}(\log N)}{N^2}\right) \to 0$ as $N \to \infty$ for fixed K so also satisfies Asymptotic Independence Theorem
- Our proposed inference procedure: hill climbing initialized using spectral clustering
  - Scales to networks with thousands of nodes and hundreds of thousands of events!

THE UNIVERSITY OF TOLEDO 1872

Comparison with discrete-time SBMs

- How does the block Hawkes model compare with discrete-time SBMs?
  - How sensitive are discrete-time SBMs to snapshot length?
- Prediction task: attempt to predict time to occurrence of next event
  - Use subset of Facebook wall post data (Viswanath et al. 2009) with ~ 3,500 nodes and ~ 140,000 events



- Block Hawkes model has lower prediction MSE than hidden Markov SBM regardless of snapshot length

Junuthula, R. R., Haghdan, M., **Xu, K. S.**, & Devabhaktuni, V. K. (2017). The Block Point Process Model for continuous-time event-based dynamic networks. arXiv preprint arXiv:1711.10967

THE UNIVERSITY OF TOLEDO 1872

Summary

- Dynamic social networks are a rich data source to learn about and predict human behavior
- Statistical models for static networks are insufficient for analyzing dynamic networks
- Proposed 2 models for discrete-time dynamic networks based on stochastic block model (SBM)
  - Hidden Markov SBM
  - Stochastic block transition model
- Proposed block point process model for continuous-time dynamic networks
  - Enables fine-resolution analysis of relational event data without need to aggregate into snapshots

THE UNIVERSITY OF TOLEDO 1872

Future Work

- Theoretical
  - Translate conditions required conditions on SBM parameters for theoretical guarantees to required conditions on point process parameters
  - Relationships between discrete-time and continuous-time dynamic network models
- Computational
  - Scaling to extremely large networks (tens or hundreds of thousands of nodes) using stochastic inference and GPU computation
  - Joint inference with discrete-time and continuous-time network data
  - More flexible models while keeping computational demands manageable

IDEAS Lab Software

https://github.com/IdeasLabUT

- Dynamic Stochastic Block Models MATLAB Toolbox (targeted to researchers)
  - Implementations of HM-SBM and SBTM for discrete-time dynamic networks
- DyNetworkX Python package (targeted to a general audience)
  - Data structures for importing and analyzing discrete-time and continuous-time dynamic networks
  - Fork of well-known NetworkX Python package (5,000+ commits from 200+ contributors)
  - Active development by 2 MS students (M. Arastuie & B. O'Leary) and 1 BS student (M. Sloma)

THE UNIVERSITY OF TOLEDO 1872