Lecture VII – Networks

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

CSE 547/STAT 548
Winter 2022

- **Connectivity**
- **Finding communities** (graph clustering)
  - Spectral clustering
- **Centrality, prestige, and authority** The goal is to give each node a score that represents its prestige or social importance. For example
  - authority of sources of information (like in PageRank or HITS) on the internet
  - impact in citation networks
  - influence, i.e. capacity of influencing others, or of attracting followers, in social networks
- **Semisupervised learning**
- **Visualization** e.g. by node embedding

Connectivity and communities

- **Connectivity**
  - Wanted large subsets of nodes that are almost disconnected from the rest.
  - can we cut the graph into two parts of comparable sizes, that have very few edges crossing between them?
- **Community detection**
  - Amounts to graph clustering
  - in computer science, social sciences, statistics, mathematics (Area where most statistical models have been developed.)
  - The quality measure for a community is called **conductance** and is related to the Normalized Cut.

$$\phi(S) = \frac{Cut(S, V \setminus S)}{\min(Vol(S), Vol(V \setminus S))} \tag{1}$$

$$NCut(S) = Cut(S, V \setminus S)\left(\frac{1}{Vol(S)} + \frac{1}{Vol(V \setminus S)}\right) \tag{2}$$

- real networks: community sizes do not grow in proportion to graph size! Hence realistic models have $K \to \infty$ when $n \to \infty$.
- Extensions: overlapping communities, nodes with features

Centrality, Influence, Authority

Various scores have been developed to quantify the above

(well understood measures)

- node degree (number of neighbors)
- eigenvector centrality
- **PageRank** and **Personalized PageRank (PPR)**

(not so well understood, may behave in unpredictable ways)

- closeness centrality

$$C_C(i) = \frac{n-1}{\sum_j d(i,j)} \tag{3}$$

- betweeness centrality

$$C_B(i) = \sum_{j,k} \frac{\sigma_{jk}(i)}{\sigma_{jk}} \tag{4}$$

where $d(i,j)$ is the graph distance and $\sigma_{jk}(i)$ is the number of shortest paths between j,k that pass through i.

- ...and many more

Semisupervised Learning

We want to estimate a function $y(i)$, $i \in V$ on the graph. For some nodes $i \in S$, y is observed; in other words these nodes are labeled, while the remaining nodes in $V \setminus S$ are unlabeled. This problem is similar to supervised learning, with the difference that we know for which future data we need to predict y.

Models for networks

- Erdos-Renyi (the null model)
- $p1$ and $p2$ models (GLM models)
- SBM (Stochastic Block Model)
- ERGM

- Latent space model
- Mixed membership SBM
- Multiplicative attributes model (overlapping communities)
- Graphons

SBM, MMSBM, and MAGM model communities explicitly.

ERGM Definitions

- $\mathcal{G} = (V, E)$ undirected graph, with $|V| = n$ nodes and edge set $E = \{ ij,\ i \neq j\} \subset V \times V$.
- **random graph model** is a distribution $P(E|V)$ defined for all finite sets of nodes V.
- equivalently, associate an indicator variable Y_{ij} to each pair of nodes, write P as a distribution of $Y|V$ with some parameters θ.
- **Exponential Random Graph Model (ERGM)** is an exponential family model for $Y_N = [Y_{ij}]_{1 \leq i < j \leq n}$.

$$P_\theta(Y_N) \ = \ \exp\left(\theta^T t_N(Y_N) - \psi_N(\theta)\right) \tag{5}$$

  - $N = n(n-1)/2$ is the dimension of Y
  - $t_N \in \mathbb{R}^d$ is a vector of sufficient statistics computed from Y.
  - dependence of V is implicit, through the dependence of N on n.

Extensions

1. Directed graphs
2. restricting the possible edges ($N = \dim Y$)
3. Considering nodes with features X_i, $i = 1 : n$ which can influence the probabilities of the edges.

$$P_\theta(Y_N|X_{1:n}) \ = \ \exp\left(\theta^T t_N(Y_N|X_{1:n}) - \psi_N(\theta, X_{1:n})\right) \tag{6}$$

The Erdös-Renyi model

Example (The Erdös-Renyi (ER) model)

For this model, $t_N = \sum y_{ij}$ and there is a single parameter $\theta \in \mathbb{R}$. Thus,

$$P_\theta(y_N) \propto e^{\theta \sum y_{ij}} = \prod_{ij} e^{\theta y_{ij}} \tag{7}$$

- Each edge is sampled iid from a Bernoulli with natural parameter θ.
- The most probable graph is the complete graph if $\theta > 0$ and the empty graph if $\theta < 0$.

The Stochastic Block-Model (SBM)

Example (The Stochastic Block-Model (SBM))

The assumption is that the nodes in V are partitioned into K clusters; $X_i \in \{1 : K\}$ denotes the cluster that i belongs to. We have $K(K+1)/2$ sufficient statistics, defined as

$$t_{kl}(y,x) = \sum_{x_i=k, x_j=l \text{ or } x_i=l, x_j=k} y_{ij} \quad (8)$$

- an edge Y_{ij} is sampled independently with a probability that dependins on where its endpoints lie.
- for known X, the normalization constant for the SBM is tractable.

- The ER and the SBM are called **diadic** models, which means that edges are sampled independently conditioned on the features of their endpoints.
- Diadic models do not fit well the real world social-networks. In particular, features like triangles and stars have higher frequency in real networks than the frequencies predicted by independent sampling of edges.

ERGM with higher order features

The sufficient statistics count other "interesting" features, like triangles, nodes of degree $k = 2, 3, 4 \ldots$, 4 and 5 cliques, in addition to edges.

Example (ERGM with star and triangle features)

Let $t_{1,N}$ count the number of edges, $t_{2,N}$ the number of triangles, $t_{3,N}$ the number of 3-stars (nodes of degree 3), $t_{4,N}$ the number of 4-stars, etc. There is a parameter θ_k for each statistic $t_{k,N}$; when $\theta_k > 0$ the model favors the graphs which contain more of feature k, and when $\theta_k < 0$ then graphs containing fewer of this feature will be more probable.

$$P_\theta(y_N) \;=\; e^{\theta_1 \#\text{edges} + \theta_2 \#\text{triangles} + \theta_3 \#\text{3-stars} + \ldots - \psi_N(\theta_1, \theta_2, \ldots)} \tag{9}$$

- these statistics will be dependent on each other
- the normalization constant Z is generally intractable.

Challenges – Algorithmic

Parameter estimation

- Estimation of parameters from a single network
- the Y variables are dependent: estimation from non-iid data.
- Sometimes the features X are dependent and not observed (e.g. SBM)
    - assume X known (easier)
    - estimate X e.g. by spectral clustering, then $\theta|X$
    - MAP/Monte Carlo estimation of both θ and X

Computational issues

- For most proper ERGMs, ψ is not computable in closed form or tractably.
- Hence sampling form P_θ and exact inferences also intractable
- For example, $P_\theta(Y_{ij} = 1|n)$ is intractable in model (9).
- typically inference by MCMC

How do we use network models?

- Model interpretation
  - predict various properties for other networks from the same source, with different n
  - e.g. number of triangles, diameter, expected degree of a node, number of edges
  - scientific interpretations
- Testing
  - does network $\mathcal{G}$ fit model P_θ ?
  - are two networks from the same source?
    Examples
  - in SBM, the expected degree grows (approximately) linearly with n – not realistic for e.g. people
  - community sizes do not grow linearly with n
  - expected number of triangles in a social network
- Parameter interpretation
  - parameter consistency – not always true!
  - independence of n

Instability and its consequences

- Assume w.l.o.g. that $t_n \in \{0, \dots T_N\}$
  - For example the number of edges $t_1 \leq N = n(n-1)/2$, the number of triangles $t_2 \leq n(n-1)(n-2)/6$, the number of 3-stars $t_3 \leq n(n-1)(n-2)(n-3)/24$.
- A sufficient statistic t_N is called **stable** iff $\frac{T_N}{N}$ is bounded as $N \to \infty$ otherwise t_n is **unstable**.
- For example, t_1 stable, t_2, t_3 unstable

Theorem (After Schweinberger)

Assume P_θ is a single parameter model with sufficient statistic t_N unstable.

1. *Denote $y_N \sim y'_N$ if the two random graphs represented by y_N, y'_N differ in the value of a single Y_{ij}. Then*
$$\max_{y_N \sim y'_N} \frac{P_\theta(y_N)}{P_\theta(y'_N)} \text{ tends to infinity when } N \to \infty.$$
(In other words, P_θ is sensitive to small changes in Y.)
2. *The probability distribution P_θ concentrates on extreme values of the sufficient statistic, i.e.*
   - *for any θ and any $\epsilon \in (0,1)$, $P_\theta[t_N(Y) \geq (1-\epsilon)T_N] \to 1$, if $\theta > 0$*
   - *or $P_\theta[t_N(Y) \leq \epsilon T_N] \to 1$, if $\theta < 0$, when $N \to \infty$.*

(In)consistency of ERGM

Definition

The sufficient statistic t **has separable increments** iff for all set of nodes B, for all $A \subset B$, and for all networks y_A, the range of possible increments $\delta = t_B(y_B) - t_A(y_A)$ is the same, and the conditional volume factor does not depend on y_A, i.e. $v_{B\setminus A|A}(\delta, y_A)$ depends only on δ.

Theorem ([Shalizi, Rinaldo,2013])

The exponential family P_θ is ***projective*** *iff the sufficient statistics have separable increments.*

For example, when a set of nodes A, with a network y_A on them, is increased with $B \setminus A$, the number of edges in examples 1, and 2, will increase by amounts that depend only on properties of A and B, but not on what edges appear in y_A. However, the number of triangles in $B \setminus A$ will depend on the configuration of edges in y_A, and in particular on the number of triangles in y_A. Hence, diadic models are projective, but ERGMs (that count triangles and stars) are not.

- Why does this matter?
- Wanted: when n increases, parameters θ must have the same meaning = projectivity