Lecture 8 - Centrality

Monday, February 7, 2022 6:48 AM

Plan for today
- Project talk ✓
- Review last week ✓
- Centrality definition ✓
- Degree, Closeness, Betweenness, Eigenvector ✓
- Back to diffusion ✓
- Epidemiology, diffusion, and centrality

Talkin' project

- Teams due in Submitty before class Monday
- Plan to present on Monday
- 10 minute max time
- Max 4 per team

What to include:
- Only a presentation

Content:
- What problem is being considered?
  → relation to graph mining

→ relation to grap

- What data to use?
  → be specific
- What algorithms/code to implement?
- What are expected outcomes?
- How will work be distributed?

→ Timeline (tentative)
  → End of March for update
  → End of April for final

Quick review from last week

Recommender systems
→ make recommendations to user
→ Graph Mining: we recommend links (predict network growth)

→ Link prediction via growth properties
→ Triadic closure, strong/weak ties
→ Generally unsupervised

→ Generally unsupervised

Matrix factorization
→ supervised predictor for links
→ learn why/why not links exist via our latent features

Collaborative filter
→ Use known user preferences to filter recommendations
→ Again, used matrix factorization

Centrality

Centrality is a measure of "importance" for a vertex
Note: related to "influence" or how a vertex might influence a diffusive

Who has high centrality?
Social networks: popular or important people

. .

important people

Road/infrastructure: important intersections

Epidemiology: someone with a large number of connections

OR small distance to everyone

Recall: hubs from first class

Degree Centrality

Defined as number of connections

- # of incident edges aka $d(v)$
- $|N(v)|$ size of neighborhood
- For directed graphs: we can consider $\underset{\text{out degree}}{d^+(v)} + \underset{\text{in degree}}{d^-(v)}$ or one or the other

Pros: - Easy to calculate

- # of connections means quick propagation of diffusive process

- # of connections [illegible]
  propagation of diffusive process
- Useful in various contexts

Cons: - Easy to falsify
- Doesn't say much for how a diffusive process will spread in network as a whole
  → very local measurement

Closeness Centrality

How "close" a vertex is to all other vertices in a network
- E.g. avg shortest path length for v
- Think in terms of how "close" a vertex is to a graph's <u>center</u>
  → How many hops/timesteps to spread a diffusive process to the rest of the network

Pros: Loosely determines how quickly

Pros: Loosely determines how quickly information might spread to rest of network from some v

Cons: Computationally complex $O(n^2)$

Betweenness centrality

The proportion of information that flows through some vertex

- Ratio of shortest u,v-paths over all u,v that go through some vertex
- For information to flow from any u to any v, it will pass some vertex a number of times proportional to betweeness centrality

Pros: Determines key vertices for information flow or 'cutpoints'

Cons: Another $O(n^2)$

Eigenvector Centrality

Note: we'll talk more thursday (PageRank)

Basically: a vertex is important if its neighbors are important

- Consider adjacency matrix A

→ solve via $A\underset{\uparrow \atop \text{eigenvector}}{x} = \lambda x$ ← eigenvalue

Pros: - Easier to calculate than other measures above (power iteration)

- Gives good intuitive results

Cons: Tough to infer from a human perspective

Diffusive Process

We've already considered diffusion w.r.t. strong/weak ties

→ we saw how "edge centrality" impacts a diffusive process

→ we saw how edge
impacts a diffusive process

→ Now: how does vertex centrality
impact diffusion?

General: We can consider statistically measuring a diffusive process in terms of how it propagates through a networks

Basic models for diffusion

- Vertex-centric behaviors
  * v updates $S[v]$ based on $S[u]$ forall $u \in N(v)$
- Complexity of network response to diffusive process is proportional to complexity of each vertex's response
  * small change in a local are can have a large non-linear global effect

---

Epidemiology

Defined as the study of the spread of diseases. For us, we're

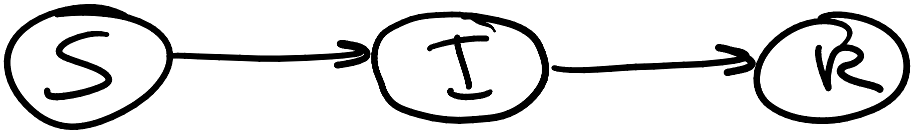
Defined as the study of the spread of diseases. For us, we're specifically concerned with impacts from network topology.

→ Can be considered a diffusive process

Here: randomness come into play

Note: while "disease" has a specific definition, the general ideas can applied to many other contexts (adoption of tech., memes, etc)

SIR model



susceptible ↑ vertex isn't infected yet

infectious ↑ vertex can transmit disease

recovered/removed ↑ vertex no longer can transmit

How our model will run:

How our model will run:

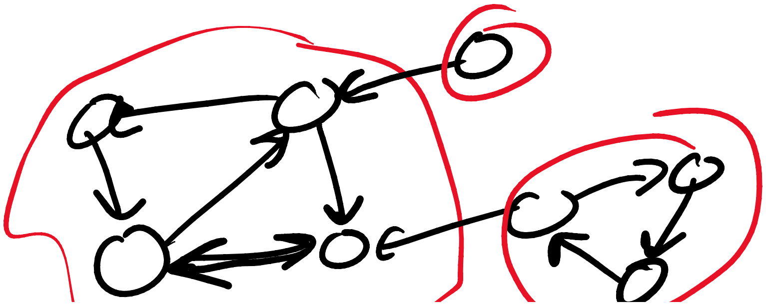
- Initialize some population as I, R
- Rest are in S
- p = probability of transmission per interaction of $S \leftrightarrow I$
- Iterate over network until $I = \emptyset$

Notes:

- p is not fixed
- interaction are not fixed
- temporal effects in network are ignored

What to watch for:

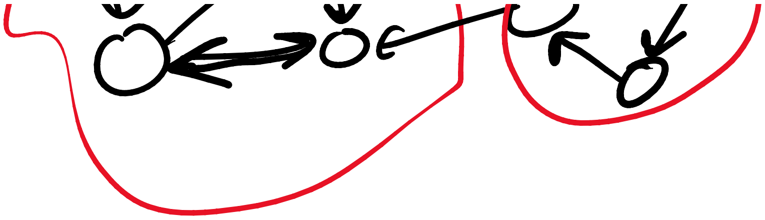
- How does centrality impact network diffusion dynamics
- Specifically in our selection of I, R initial vertex



3 total

2 nontrivial

1 trivial



1 trivial