Extracting Information from Complex Networks

1

Complex Networks

- Networks that arise from modeling complex systems: relationships
  - Social networks
  - Biological networks
- Distinguish from
  - random networks
  - uniform networks
    - grid
    - ring

2

Social networks

- Model relationships between people directly or indirectly
  - Relationships in social networking sites
    - Facebook friends
    - Twitter followers
  - Citation networks
    - twitter retweets
    - Wikipedia
    - paper citations
- Not clear separation social networks from other complex networks

3

Information from network structure

- Explore properties of graph
  - nodes
  - edges
- Interpret in context of subject of network

4

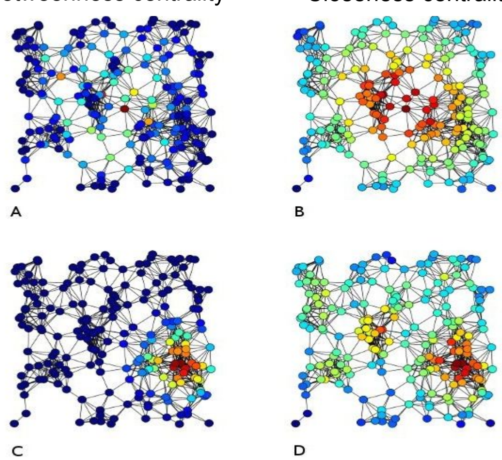
Graph measures of interest for nodes

- degree/indegree/outdegree centrality
- pagerank
- sum of distances to all other nodes
  - Reciprocal is closeness centrality
- betweenness centrality
  - measure based on number of shortest paths in graph that go through the node
- cluster coefficient
  - fraction of pairs of neighbors of node that have edge between them

Uses

- Look at nodes that stand out under different measures
- Look at distribution of values of measure

Betweenness centrality   Closeness centrality



Eigenvector centrality   Degree centrality

From Wikipedia “centrality” article

Graph properties of interest for network

- density
  (number of edge)/(number of possible edges)
    directed vs undirected? self-edges?
- diameter
  largest shortest path
- distribution of shortest paths
    “6 degrees of separation”
- average cluster coefficient
- distribution of degrees

Characterizing social networks

for social network with n nodes

- average density low
- average shortest path log(n) or less
  - small world network
- form communities
- distribution of degrees follows power law
  - power law: log(y) = a*log(x) + b
    - eg Zipf's law
  - call "scale-free"

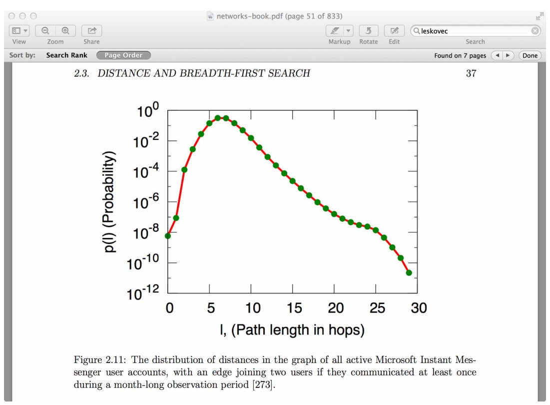
9

Small world phenomena

- Travers & Milgram 1969 *Sociometry*
  - 296 letters to start; 67 reached target person
  - Mean length path followed 6.2
- Leskovec & Horvitz 2008 *WWW Conf*
  - Microsoft Instant Messenger, 240 million active users
  - Edge: two-way conversation
  - One giant component
  - Average distance 6.6
  - 90% effective diameter 7.8

10

From *Networks, Crowds and Markets*



11

Characterizing relationships

- Relationship: edge between two nodes
  - Consider now just undirected
  - Refer to as "neighbors"
- Would like to extract properties of the relationship from network structure.
- Measures – here are two
  - Embeddedness: number of mutual neighbors
  - Dispersion: measure of connectedness among mutual neighbors
    - Backstrom & Kleinberg, 2014

12

A network Analysis of Relationship Status on Facebook

Backstrom & Kleinberg 2014

- Observe: person's network of friends represents diverse set of relationships
- Question: Can one recognize romantic partners on Facebook from structure of friends network?
- Contributions (some)
  - Define new measure dispersion
  - Show dispersion works better that embeddedness
  - Show dispersion works pretty well
  - Show combining dispersion with many other signals via machine learning does even better

Dispersion Definition

- Actually define several versions
- Basic: absolute dispersion disp(u,v) for link (u,v)
  - u distinguished: want to predict his/her partner
  - Define G_u as the subgraph on neighbors of u
  - Define $C_{u,v}$ as the set of common neighbors of u and v
  - For s,t nodes in $C_{u,v}$, define $f_{u,v}(s,t)$ with value

    1 if s, t not neighbors and have no common neighbors in G_u other than u and v

    0 otherwise
  - $\text{disp}(u,v) = \sum_{s,t \text{ in } C_{u,v}} f_{u,v}(s,t)$

Experiments: Data

- Facebook users
  - At least 20 years old
  - Between 50 and 2000 friends
  - Listed spouse or relationship partner on profile
- Sample ~1.3 million of these users selected uniformly at random and their network neighborhoods (extended dataset)
  - Neighborhoods avg 291 nodes, 6652 links
  - 379 million nodes , 8.8billion links overall
- Subsample 73,000 neighborhoods (primary dataset)
  - Only neighborhoods with at most 25,000 links
  - Uniformly at random

Experiments: Modify definition of dispersion

- For improved results
- Normalized dispersion: disp(u,v)/emb(u,v)
  - emb(u,v) is embeddedness
- Recursive dispersion: look at neighbors of neighbors of neighbors …
  - Find best performance using 3 levels

Experiments: results (some)

type	embed	rec.disp.	photo	prof.view.
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (fem)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
engaged	0.179	0.446	0.442	0.391
engaged (fem)	0.171	0.399	0.386	0.401
engaged (male)	0.185	0.490	0.495	0.381
relationship	0.132	0.344	0.347	0.441
relationship (fem)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

Figure 4. The performance of different measures for identifying spouses and romantic partners: the numbers in the table give the *precision at the first position* — the fraction of instances in which the user ranked first by the measure is in fact the true partner. Averaged over all instances, recursive dispersion performs approximately twice as well as the standard notion of embeddedness, and also better overall than measures based on profile viewing and presence in the same photo.

Additional questions in paper

- How much better can lots of features do?
  - Combined 120 features for nodes in primary dataset
    - Combined variations of dispersion def
    - Included many other properties from user pages and behavior
  - Used machine learning classifier
    - Trained on 50% users
  - Overall precision at 1st position 0.705 (vs 0.506)

Additional questions in paper

- What about predicting whether in a relationship?
  - High dispersion link from u does not mean romantic relationship
    - Property is bridging groups of u's friends
      - family, close friends
  - Used machine learning yes/no classifier
    - 68.3% accuracy single vs any relationship
      - Baseline 59.8 – predict more common class
    - 79.0% accuracy single vs married
      - Baseline 56.6
  - Max over user's friends of normalized dispersion most important of network features used

Finding Communities

Clustering

- General clustering algorithms don't work well for graphs of unweighted edges
  - Agglomerative?
  - Divisive?
- Used different techniques
  1. Betweenness based

Betweenness definition

- Gave you:

Edge Betweeenness = # shortest paths using edge

Betweenness definition

- Gave you:

~~Edge Betweeenness = # shortest paths using edge~~

Real definition:

- For an edge e:
  - for each pair of nodes x and y in the graph, e is credited with the fraction of shortest paths between x and y that contain e
  - Sum credits over all n(n-2)/2 pairs x,y

Using Betweenness in Community Finding

- Repeat until graph disconnected:
  - Remove edge with largest betweenness
  - Recalculate betweenness
- Graph can fall into one or more pieces
- Can repeat on pieces until find desired number or size of communities =>

  Hierarchical divisive

How calculate betweenness Girvan-Newman Algoritm

- Repeat for each node x in graph the following 2 steps:

  1. Do breadth first search from node x
     - Induces parent/child relationship
     - As search, label each node with number of shortest paths from x to it:
       - Level by level: sum of labels of parents
       - Include x to itself (1)

2. Working bottom up, level by level, calculate credits for each node and then credits for edges from level above:
   - Each leaf gets 1 credit
   - Calculate edge credits for edges to level above
   - Calculate node credits for next level up

   - Each non-leaf gets 1 credit plus sum of credits on edges from it to next level below.
     - Edge credits already calculated.

Edge credit

- Given a node b, not the root, let b have parents $a=a_1, a_2, \ldots a_k$
- Let parent a_i have p_i shorteset paths to it
  - Calculated step one
- Credit for edge (a_i,b)=

$$((\text{credit for b})*p_i) / \Sigma^k_{j=1} p_j$$

Final calculation

- Now have n edge credits per edge - one for breadth first search starting at each node as root.
- Sum the n credits for an edge.
- Divide by two for final edge betweenness
  - Double-counted paths

Using Spectral Clustering

- Goal: bi-partition undirected graph
  - want each partition of close to equal size
- Define diagonal matrix D:
  - D(i,i) = degree of node i
- Define Laplacian matrix L = D-E
  - for adjacency matrix E
- Look at 2^{nd} smallest eigenvalue of L

29

Specifics

- smallest eigenvalue of L = 0
  - eigenvector all 1's (denote as **1**)
- second smallest eigenvalue:

  minimize $\mathbf{x}^T L \mathbf{x}$ such that

  X orthogonal to **1** (i.e. $\Sigma_i x_i = 0$)

  X unit vector (i.e. $\Sigma_i x_i^2 = 1$)
- show equivalent to

  minimize $\Sigma_{\text{edges }(i,j)} (x_i - x_j)^2$

  under same constraints

30

Partitioning

- x_i's must be positive and negative
  - $\Sigma_i x_i = 0$ and $\Sigma_i x_i^2 = 1$
- Nodes with positive x_i s in one partition
- Nodes with negative x_i s in other
- Properties
  - minimization tends to give x_i, x_j same sign when is edge (i,j) => minimizing cut
    - minimizing $\Sigma_{\text{edges }(i,j)} (x_i - x_j)^2$
  - minimization tends to balance sizes

31

HITS and clustering

Recall HITS matrix formulation:

$\boldsymbol{a} = E^T\boldsymbol{h}$ $\qquad$ $\boldsymbol{a} = E^TE\boldsymbol{a}$

$\boldsymbol{h} = E\boldsymbol{a}$ $\qquad$ $\boldsymbol{h} = EE^T\boldsymbol{h}$

for adjacency matrix E, authority vector **a**, hub vector **h**

- ***a*** is the eigenvector corresponding to the eigenvalue 1 for E^TE
- ***h*** is the eigenvector corresponding to the eigenvalue 1 for EE^T

32

HITS and clustering

- Non-principal eigenvectors of EE^T and E^TE have positive and negative component values
  - Denote $a_{e2}, a_{e3}, \dots$
    matching $h_{e2}, h_{e3}, \dots$
- For a matched pair of eigenvectors $\boldsymbol{a}_{ej}$ and $\boldsymbol{h}_{ej}$
  - Denote k^{th} component of j^{th} pair: $\boldsymbol{a}_{ej}(k)$ and $\boldsymbol{h}_{ej}(k)$
  - Make a "community" of size *c* (chosen constant):
    - Choose c pages with most positive $\boldsymbol{h}_{ej}(k)$ - hubs
    - Choose c pages with most positive $\boldsymbol{a}_{ej}(k)$ - authorities
  - Make another "community" of size *c*:
    - Choose c pages with most negative $\boldsymbol{h}_{ej}(k)$ - hubs
    - Choose c pages with most negative $\boldsymbol{a}_{ej}(k)$ - authorities

33

Do all social networks, as networks, have same properties?

- Kwak, Lee, Park, Moon study Twitter
  (pub 2010):

  NO

34

Kwak, Lee, Park, Moon experimental set-up

- July 6-31, 2009 crawl of Twitter
  - 41.7 million user profiles collected
  - 1.47 billion social relations
- started with "Paris Hilton" and crawled followers and "followings"
- Added users tweeting about trending topics
  - 4,262 trending topics
    - collected top ten every 5 minutes
  - 106 million tweets mentioning trending topics

35

Kwak, Lee, Park, Moon Findings

- # followers fits power law but
- users with > 100,000 followers have many more followers than expect
- 77.9% links one way
- shortest path between users shorter than other social networks
  - average 4.12
  - for 97.6 % pairs, path length ≤ 6

36

Kwak, Lee, Park, Moon: ranking users

- followers graph
  - – number of followers
  - – PageRank

  similar rankings
- retweets of user's posts
  - – very different from graph measures

37

Summary: Complex Networks and Obtaining Information

- Complex networks provide many ways of improving our acquisition of information
- Uses still in active development

38