SNA Applications I

CS 224W
Why are you taking SNA?

- Your project will hopefully get you (at least) part way there
  - You want to understand aspects of a real-world network
  - You want to understand underlying mechanisms shaping the networks around us
  - You want to develop new algorithms that take advantage of the network features in the data
What makes a good SNA project?

- Correlating network measurement with relevant (non-network) variables
- Assumptions about network formation yield predictions of observed structure
- Network features outperform others in prediction task
Tie strength & information
Are strong ties “local”? 

- A strong tie
  - frequent contact
  - affinity
  - many mutual contacts

“forbidden triad”: strong ties are likely to “close”
M. S. Granovetter: *The Strength of Weak Ties*, AJS, 1973:

- finding a job through a contact that one saw
  - frequently (2+ times/week) 16.7%
  - occasionally (more than once a year but < 2x week) 55.6%
  - rarely 27.8%

- but... length of path is short
  - contact directly works for/is the employer
  - or is connected directly to employer
Strong ties are less likely to be a bridge (or a local bridge)

edge embeddeness

- embeddeness: number of common neighbors the two endpoints have

- neighborhood overlap:

  \[
  \frac{\text{number of nodes who are neighbors of both } A \text{ and } B}{\text{number of nodes who are neighbors of at least one of } A \text{ or } B}
  \]
snowball sampling:

will you reach more different kids by asking each kid to name their 2 best friends, or their 7th & 8th closest friend?

the strength of intermediate ties

- **strong ties**
  - frequent communication, but ties are redundant due to high clustering

- **weak ties**
  - reach far across network, but communication is infrequent...

- Onnela J. et.al. PNAS 2007;104:7332-7336
  - use nation-wide cellphone call records and simulate diffusion using actual call timing
  - in simulation, individuals are most likely to obtain novel information through ties of intermediate strength
Characterizing the large-scale structure and the tie strengths of the mobile call graph

Onnela J et al. PNAS 2007;104:7332-7336

©2007 by National Academy of Sciences
Edge neighborhood overlap as a function of tie strength
The dynamics of spreading on the weighted mobile call graph, assuming that the probability for a node $v_i$ to pass on the information to its neighbor $v_j$ in one time step is given by $P_{ij} = xw_{ij}$, with $x = 2.59 \times 10^{-4}$.
is it good to be embedded?

- What are the advantages of occupying an embedded position in the network?
- What are the disadvantages of being embedded?
- Advantages of being a broker (spanning structural holes)?
- Disadvantages of being a broker?
Managers asked to come up with an idea to improve the supply chain

Then asked:
- whom did you discuss the idea with?
- whom do you discuss supply-chain issues with in general
- do those contacts discuss ideas with one another?

673 managers (455 (68%) completed the survey)
~ 4000 relationships (edges)
Figure 2. Supply-Chain Discussion Network (excludes 193 social isolates)

Source: Structural Holes and Good Ideas; R. Burt, American Journal of Sociology, 2004
Network Constraint

\[ C = \sum_{ij} c_{ij} = \sum_{ij} \left[ p_{ij} + \sum_{q} p_{iq} p_{qj} \right] \]

\[ \text{person 2: } 0.265 = \left( \frac{1}{3.5} + 0 \right)^2 + \left( \frac{1}{3.5} + 0 \right)^2 + \left( \frac{1}{3.5} + 0 \right)^2 + \left( \frac{1}{3.5} + 0 \right)^2 \]

\[ \text{person 3: } 0.402 = \left( \frac{1}{0.25} + 0 \right)^2 + \left( \frac{1}{0.25} + 0.084 \right)^2 + \left( \frac{1}{0.25} + 0.091 \right)^2 + \left( \frac{1}{0.25} + 0.084 \right)^2 \]

\[ \text{Robert: } 0.148 = \left( \frac{1}{0.077} + 0 \right)^2 + \left( \frac{1}{0.154} + 0 \right)^2 + \left( \frac{1}{0.154} + 0 \right)^2 + \left( \frac{1}{0.154} + 0 \right)^2 + \left( \frac{1}{0.154} + 0 \right)^2 + \left( \frac{1}{0.154} + 0 \right)^2 \]

Source: Structural Holes and Good Ideas; R. Burt, American Journal of Sociology, 2004
results

- people whose networks bridge structural holes have
  - higher compensation
  - positive performance evaluations
  - more promotions
  - more good ideas

- these brokers are
  - more likely to express ideas
  - less likely to have their ideas dismissed by judges
  - more likely to have their ideas evaluated as valuable
Position -> information -> $$$?

Betweenness

Constrained vs. Unconstrained


slide: Marshall van Alstyne
Study of a head hunter firm

- Three firms initially
- Unusually measurable inputs and outputs
  - 1300 projects over 5 yrs and
  - 125,000 email messages over 10 months (avg 20% of time!)
- Metrics
  - (i) Revenues per person and per project,
  - (ii) number of completed projects,
  - (iii) duration of projects,
  - (iv) number of simultaneous projects,
  - (v) compensation per person
- Main firm 71 people in executive search (+2 firms partial data)
  - 27 Partners, 29 Consultants, 13 Research, 2 IT staff
- Four Data Sets per firm
  - 52 Question Survey (86% response rate)
  - E-Mail
  - Accounting
  - 15 Semi-structured interviews

### New Contract Revenue Coefficients

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unstandardized Coefficients</th>
<th>Adjusted R²</th>
<th>Significance F Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Base Model)</td>
<td></td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Best structural pred.</td>
<td>12604.0***</td>
<td>4454.0</td>
<td>0.52 .006</td>
</tr>
<tr>
<td>Ave. E-Mail Size</td>
<td>-10.7**</td>
<td>4.9</td>
<td>0.56 .042</td>
</tr>
<tr>
<td>Colleagues’ Ave. Response Time</td>
<td>-198947.0</td>
<td>168968.0</td>
<td>0.56 .248</td>
</tr>
</tbody>
</table>

### Contract Execution Revenue Coefficients

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unstandardized Coefficients</th>
<th>Adjusted R²</th>
<th>Significance F Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Base Model)</td>
<td></td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>Best structural pred.</td>
<td>1544.0**</td>
<td>639.0</td>
<td>0.30 .021</td>
</tr>
<tr>
<td>Ave. E-Mail Size</td>
<td>-9.3*</td>
<td>4.7</td>
<td>0.34 .095</td>
</tr>
<tr>
<td>Colleagues’ Ave. Response Time</td>
<td>-368924.0**</td>
<td>157789.0</td>
<td>0.42 .026</td>
</tr>
</tbody>
</table>

---

**Sending shorter e-mail helps get contracts and finish them.**

**Faster response from colleagues helps finish them.**

---

Email structure

AVERAGE TIME SPENT COMPOSING ONE E-MAIL

PROFESSORS: 1.3 SECONDS

YES.  (SEND)

DO IT.  (SEND)

SEE ATTACHED.  (SEND)

NO.  (SEND)

GRAD STUDENTS: 1.3 DAYS

DEAR (?) PROF. SMITH,

I WAS WONDERING IF PERHAPS YOU MIGHT HAVE POSSIBLY GOTTEN THE CHANCE TO POTENTIALLY FIND THE TIME TO MAYBE LOOK AT THE DRAFT PAPER THAT I AM ATTACHING (IN JUST IN CASE). IF YOU ARE NOT, THEN IF YOU HAVE ANY QUESTIONS, WHATSOEVER, PLEASE DO NOT HESITATE TO...

WWW.PHDCOMICS.COM
### H5: Recruiters with larger personal rolodexes generate no more or less output

<table>
<thead>
<tr>
<th></th>
<th>Revenue $</th>
<th>$ for completed searches</th>
<th>Completed searches</th>
<th>Multitasking</th>
<th>Duration</th>
<th>Duration controlling for multitasking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size of rolodex (Q50)</td>
<td>-10.2</td>
<td>-22.9</td>
<td>0.000</td>
<td>0.000</td>
<td>-0.013</td>
<td>-0.013</td>
</tr>
<tr>
<td></td>
<td>(60.3)</td>
<td>(32.6)</td>
<td>(0.001)</td>
<td>(0.001)</td>
<td>(0.021)</td>
<td>(0.016)</td>
</tr>
</tbody>
</table>

* p < 0.10, ** p < 0.05, *** p < 0.01, Standard err in paren.

Instead, a larger private rolodex is associated with:

- Less information sharing
- Less DB proficiency
- Lower % of e-mail read
- Less learning from others
- Less perceived credit for ideas given to colleagues
- More dissembling on the phone

Information access

- network structure (having high degree) correlates with receiving novel information sooner (as deduced from hashed versions of their email)

- getting information sooner correlates with $$ brought in
  - controlling for # of years worked
  - job level
  - ....

Strong ties may be more redundant, but they also share information more frequently, and are willing to share novel information.

Bridging diverse communities is significant.

Being in the thick of information flows is significant.

Information backbone

Kossinets, Watts, Kleinberg, KDD 2008:
- which paths yield the most up to date info?
- how many of the edges form the “backbone”?

image source: Kossinets et al. “The structure of information pathways in a social communication network”, KDD 2008
Motifs
Resolving local structure: network motifs

motif matches in the target graph

http://mavisto.ipk-gatersleben.de/frequency_concepts.html
All 3 node motifs
Examples of network motifs (3 nodes)

- **Feed forward loop**
  - Found in neural networks
  - Seems to be used to neutralize “biological noise”

- **Single-Input Module**
  - e.g. gene control networks
Examples of network motifs (4 nodes)

- Parallel paths
  - Found in neural networks
  - Food webs
4 node subgraphs (computational expense increases with the size of the graph!)
Compare to “equivalent” random graph

Some motifs will occur more often in real world networks than random networks.

**Technique:**
- construct many random graphs with the same number of nodes and edges (same node degree distribution?)
- count the number of motifs in those graphs
- calculate the Z score: the probability that the given number of motifs in the real world network could have occurred by chance

**Software available:**
- [http://www.weizmann.ac.il/mcb/UriAlon/](http://www.weizmann.ac.il/mcb/UriAlon/) (the original)
- FanMod [http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html](http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html) (faster and more user friendly)
What the Z score means

\[ Z_X = \frac{X - \mu_X}{\sigma_X} \]

\[ \mu = \text{mean number of times the motif appeared in the random graph} \]

\[ \sigma = \text{standard deviation} \]

The probability of observing a Z score of 2 is 0.02275.

In the context of motifs:
- \( Z > 0 \), motif occurs more often than for random graphs.
- \( Z < 0 \), motif occurs less often than in random graphs.
- \(|Z| > 1.65\), only a 5% chance of random occurrence.
software: FANMOD (also igraph)

http://theinf1.informatik.uni-jena.de/~wernicke/motifs/index.html

FANMOD a tool for fast network motif detection
Superfamilies of networks

source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004
Quiz Q:

Based on their triad census profiles, which two kinds of networks exhibit similar structure?

(a) Transcription and language
(b) WWW and language
(c) Social and WWW
(d) Language and social
Superfamilies of networks

source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004
Quiz Q:

Which of the following triads is underrepresented in social networks?

(a) 6    (b) 9    (c) 12    (d) 13
Superfamilies of networks

source: Milo et al., Superfamilies of Evolved and Designed Networks, Science 303:1538-1542, 2004
Motifs: recap

- Given a particular structure, search for it in the network, e.g. complete triads

- Advantage: motifs can correspond to particular functions, e.g. in biological networks

- Disadvantage: don’t know if motif is part of a larger cohesive community
Modeling community structure in a Q&A forum

Zhang, Ackerman, Adamic, WWW’07
simulating probability of expertise pairing

suppose:

- expertise is uniformly distributed
- probability of posing a question is inversely proportional to expertise

\[ p_{ij} = \text{probability a user with expertise } j \text{ replies to a user with expertise } i \]

2 models:

\[ p_{ij} \sim i^{-1} e^{\beta (j-i)} \quad \text{‘best’ preferred} \]

\[ p_{ij} \sim i^{-1} e^{\gamma (i-j)} \quad j > i \]
Best “preferred”  just better
degree correlation profiles

degree-degree correlations between asker and helper (empirically observed)

best preferred (simulation)  just better (simulation)
But motifs reveal that we are not capturing all the local structure!
Egonetworks: another way of capturing local structure

(a) Programming
(b) Marriage
(c) Wrestling

Adamic, Zhang, Bakshy, Ackerman WWW'08
Recipe recommendation using ingredient networks

Chun-Yuen Teng\textsuperscript{1}, Yu-Ru Lin\textsuperscript{2,3}, Lada A. Adamic\textsuperscript{1}

\textsuperscript{1}School of Information, University of Michigan
\textsuperscript{2}IQSS, Harvard University
\textsuperscript{3}CCS, Northeastern University

WebSci’12
Crispy Herb Baked Chicken

By: DCANTER
"The secret ingredient is instant mashed potatoes, used to make the crispy coating."

Rate/Review | Read Reviews (438)

Add a photo
1 of 23 Photos

Prep Time: 15 Min  Cook Time: 45 Min  Ready In: 1 HR

Servings (Help)

4  US  Metric  Calculate

Original Recipe Yield 4 - 5 servings

Ingredients

2/3 cup dry potato flakes
1/3 cup grated Parmesan cheese
1 teaspoon garlic salt
1 (3 pound) chicken, skin removed, cut into pieces
1/3 cup butter, melted

Directions
Our online recipes

- allrecipes.com
- 46,337 recipes
- Each recipe includes directions, ingredients, nutrition info, cooking time, and regional information
- 1,976,920 reviews include ratings and text
Research questions

- What patterns emerge from the collective cooking knowledge aggregated in recipes?
- How can ingredient networks be used for predicting recipe ratings?
Recipe mining

- Cooking methods
  - Regional preferences
- Ingredients
  - Combination of ingredients
  - Modification of ingredients
- Predicting ratings:

Credit: sonomaorganics.com
Ingredients
Combining ingredients

- Banana + Blueberries = ❖
- Apple + Corn = ❌
Complement network

- Nodes: ingredients

- Undirected edges:
  - Weighted by pointwise mutual information
    \[ \text{PMI} = \log \left( \frac{P(a,b)}{P(a)P(b)} \right) \]
  - \[ P(a,b) = \frac{\text{(\# of recipes containing a and b)}}{\text{(\# of recipes)}} \]
  - \[ P(a) = \frac{\text{(\# of recipes containing a)}}{\text{(\# of recipes)}} \]
  - \[ P(b) = \frac{\text{(\# of recipes containing b)}}{\text{(\# of recipes)}} \]

- Recipe rating and PMI of its ingredient pairs
  - Mean and minimum of PMI (no correlation with rating)
  - Max of PMI (\(\rho=0.09, p < 0.001\))
Complement network
With 3 clusters of ingredients

sweet

mixed
drinks

savory

olive oil
garlic
parsley
black pepper
basil
chicken
broth
tomato
celery
pepper
onion

butter
white sugar
flour
vanilla extract
brown sugar
milk

baking powder
confectioners' sugar
walnut
honey

water
vegetable oil

marsala cherry
pineapple juice
Guess the cuisine!

What do you notice about sweet vs. savory ingredients?
Recipe modification

- Egg times 2
- Egg times 3

- Strawberries
- Orange juice
- Blueberries
Recipe modification

• 60% of reviews contain “add”, “omit”, “instead”, “extra”, and 14 others.

• Reviews that include changes assign higher star ratings (4.49 vs. 4.39, \(p<10^{-10}\))

• almost perfect but not quite (4 star) reviews often suggest modifications
Suggested modifications of quantities

![Graph showing suggested modifications of quantities with various ingredients plotted on a scatter plot. The x-axis represents the number of reviews adjusting down per recipe, and the y-axis represents the number of reviews adjusting up per recipe. Ingredients such as salt, butter, egg, flour, sugar, water, onion, garlic, milk, vanilla, pepper, olive oil, brown sugar, black pepper, cinnamon, tomato, pepper, olive oil, and more are plotted on the graph.]
Correlations between ingredient modifications

- Recipe freq. vs. deletion/recipe ($\rho = -0.22$)
- Recipe freq. vs. addition/recipe ($\rho = -0.25$)
- Recipe freq. vs. increase/recipe ($\rho = -0.26$)

<table>
<thead>
<tr>
<th></th>
<th>addition</th>
<th>deletion</th>
<th>increase</th>
<th>decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td># recipes</td>
<td>0.41</td>
<td>0.22</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>addition</td>
<td></td>
<td>-0.15</td>
<td>0.79</td>
<td>0.11</td>
</tr>
<tr>
<td>deletion</td>
<td></td>
<td></td>
<td>0.09</td>
<td>0.58</td>
</tr>
<tr>
<td>increase</td>
<td></td>
<td></td>
<td></td>
<td>0.39</td>
</tr>
</tbody>
</table>

Recipe freq. vs. deletion/recipe ($\rho = -0.22$)
Recipe freq. vs. addition/recipe ($\rho = -0.25$)
Recipe freq. vs. increase/recipe ($\rho = -0.26$)
Substitution network

- Extract substitution relationships from comments
  - e.g. “I replaced the butter with sour cream”
  - “replace a with b”, “substitute b for a”, “b instead of a”

- Nodes: ingredients

- Edge weights = $p(b | a)$, which is the proportion of substitutions of ingredient a that suggests ingredient b
Try this at home: maapeakequation.org
Substitution network: communities
Preference network

- Create an edge from ingredient \( a \) to \( b \) if \( \text{rating}(a) < \text{rating}(b) \)

- ex:
  - Recipe X contains grapes
  - Recipe Y contains ketchup
  - Rating(X) > Rating(Y)
Substitute network and users’ preference

- Weight of preference network
  - PMI(a→b) = \log(p(a→b)/p(a)p(b))
  - where p(a→b) = (# of recipe pairs from a to b)/(# of recipe pairs)

- Correlations between preference network and substitute network (\rho = 0.72, p<0.001)
Prediction task

Given a recipe pair with overlapped ingredients, determine which one has the higher rating.
Prediction task

- **Features**
  - **Baseline**
    - Cooking methods, preparation time, the number of servings
  - 1000 popular ingredient list
    - Binary vector indicating the occurrence of ingredients
  - **Nutrition**
    - Calories, carbohydrates, fat, etc.
  - **Ingredient networks**
    - Network positions (centrality) and communities (SVD)
  - **Combined set**
    - Everything listed above
Prediction task

- 62,031 recipe pairs \((X,Y)\)
  - where \(\text{rating}(X) > \text{rating}(Y)\)
  - \(\geq 10\) user reviews
  - \(\geq 50\%\) users have rated both recipes
  - Cosine similarity of ingredients \((X,Y) > 0.2\)

- Train with gradient boosting tree
  - balanced dataset
  - 2/3 for training, 1/3 for testing
  - Evaluate based on accuracy
Ingredient network features lead to improved performance
Recipes encode our collective cooking knowledge

- complementarity of ingredients
- regional preferences in combining ingredients
- substitutability of ingredients
- complement and substitute networks encode users’ preferences and can be used to effectively predict recipe ratings
We saw diverse applications
- Tie strength
- Network diversity
- Motifs
- Centrality & community structure

What makes it interesting:
- Measurable consequence: tasty food (ratings), information (email topics), respect (survey responses)!
- Models!