Discovering Communities through Information Structure and Dynamics

Nick Montfort

Written Preliminary Examination II
Computer and Information Science
University of Pennsylvania

9 August 2004

Organizations, communities, and information

- Communities of many sorts exist within larger organizations, whether they are traditional or distributed.
- The structure of communities is important, but is almost never documented or easily determined.
- Computer communications and online documents can reveal this structure.

Why learn about communities?

To infiltrate and destroy them!
(Actually, probably not...)

To help people and organizations reach their goals; foster productivity, learning, research, creativity; support and build community.

Studies of communities and information flow/structure

Three dealing with the **World Wide Web:**

- Gibson, Kleinberg, & Raghavan 1998
- Flake, Lawrence, & Giles 2000
- Adamic & Adar 2003

Two dealing with **email data:**

- Tyler, Wilkinson, & Huberman 2003
- Eckmann, Moses, & Sergi 2003

What are “communities”?

Community was once used only to mean a local group, such a village, town, or neighborhood.

Sociologists now use the term more broadly, and not just because of computer-mediated communications.

Defining “community”

Organization: any stable pattern of transactions between individuals or aggregations of individuals.

Community: a social network of relationships that provide sociability support, information, and a sense of belonging.

Determining communities on the Web with HITS

Three dealing with the **World Wide Web:**

- Gibson, Kleinberg, & Raghavan 1998
- Flake, Lawrence, & Giles 2000
- Adamic & Adar 2003

Two dealing with **email data:**

- Tyler, Wilkinson, & Huberman 2003
- Eckmann, Moses, & Sergi 2003

Hypertext Induced Topic Search (HITS)

A way of discovering the best “hubs” and “authorities” for a given topic, starting with a topic-specific set of pages and then using only the hypertext structure.

From Kleinberg 1999; basis for the work in Gibson, Kleinberg, & Raghavan 1998.

Characterizes communities by their important pages, but doesn’t partition a graph into community/non-community.

HITS: Hubs and authorities

A **root set** (from a search result) is the starting point; this is then expanded.

Hub: a page that is good at pointing to other pages—e.g., respected directories.

Authority: a page that is good at being pointed to—e.g., reliable sources.

Hubs point to many authorities, authorities are pointed to by many hubs.

HITS: Iterative algorithm

Hub weights: h_i Authority weights: a_i

for (i=1..n)
  $a_i \leftarrow 1$; $h_i \leftarrow 1$;
for (i=1..k)
  for (p=1..n)
    $a_p \leftarrow \sum_{q:(q,p)\in E} h_q$; $h_p \leftarrow \sum_{q:(p,q)\in E} a_q$;
    a ← normalize(a); h ← normalize(h);

HITS: Eigenvector formulation

$a_p \leftarrow \sum_{q:(q,p)\in E} h_q$; $h_p \leftarrow \sum_{q:(p,q)\in E} a_q$;

With adjacency matrix A, one step:
$a \leftarrow Ah$; $h \leftarrow A^T a$;

With z a vector of all ones, k steps:
$h \leftarrow (A^TA)(A^TA)\ldots(A^TA)(A^TA)z$; that is...
$h \leftarrow (A^TA)^k z$; Similarly: $a \leftarrow (AA^T)^k z$;

So, the whole algorithm is just:
$a \leftarrow \omega_1(AA^T)$; $h \leftarrow \omega_1(A^TA)$;

HITS: Multiple, possibly overlapping communities

$\omega_1(AA^T)...\omega_n(AA^T)$ correspond to different communities, with the components of ω_i indicating the most important authorities.

Similarly for hubs and $\omega_1(A^TA)...\omega_n(A^TA)$

One page may be a hub/authority in multiple communities.

HITS and communities: Main advantages/problems

Models multiple, overlapping communities.

Reflects that Web pages can be good referrers or good pages to refer to.

Only one free parameter: The root set.

This root set is very high-dimensional and may be generated with (unknown) hackery!

Partioning a community with an approximate minimum cut

Three dealing with the **World Wide Web:**

- Gibson, Kleinberg, & Raghavan 1998
- Flake, Lawrence, & Giles 2000
- Adamic & Adar 2003

Two dealing with **email data:**

- Tyler, Wilkinson, & Huberman 2003
- Eckmann, Moses, & Sergi 2003

Approximate min cut: The idea

For any page s on the Web, find a "community": the set of pages, including s, that has more links (both ways) within the set than to pages outside the set.

On an undirected graph G, pick t not in the community. An s-t minimum cut C identifies the community, if s links to more than |C| community vertices and t links to more than |C| non-community vertices

Approximate min cut: The reality

Crawl from “seed” vertex, recrawl 3 times
Add bidirectional edges 1st gen to 2nd gen
Adjust some edge weights
Use a “virtual sink,” link to remote vertices

Three good experimental results
No proof of convergence
No proof of quality of approximation

Approximate min cut vs. HITS

Can enumerate community members without extracting multiple eigenvectors.

Doesn’t rely on identifying dominant hub/authorities, may identify communities of other topologies.

Begins with just a URL, not a root set.

Points out that the selection of a root set may be doing a lot of work.

Finding Web “friends” with text and link similarity

Three dealing with the **World Wide Web:**

- Gibson, Kleinberg, & Raghavan 1998
- Flake, Lawrence, & Giles 2000
- Adamic & Adar 2003

Two dealing with **email data:**

- Tyler, Wilkinson, & Huberman 2003
- Eckmann, Moses, & Sergi 2003

Finding “friends”: Similarity predicting links

Four types of items i on home pages A, B:
named entities, in-links, out-links, mailing list membership

Similarity(A,B) = $\sum_{i \in A, i \in B}$ 1/log(count(i))

count(i) counts the total occurrences

Which types of items, and which specific items, best predict a link between A and B?

Finding “friends”: Best predictors

1. In-links
2. Mailing list membership
3. Out-links
4. Named entities

Different items at MIT and Stanford

But, are these the right textual (and structural) measures to look at?

Clustering the email graph with betweenness

Three dealing with the **World Wide Web:**

- Gibson, Kleinberg, & Raghavan 1998
- Flake, Lawrence, & Giles 2000
- Adamic & Adar 2003

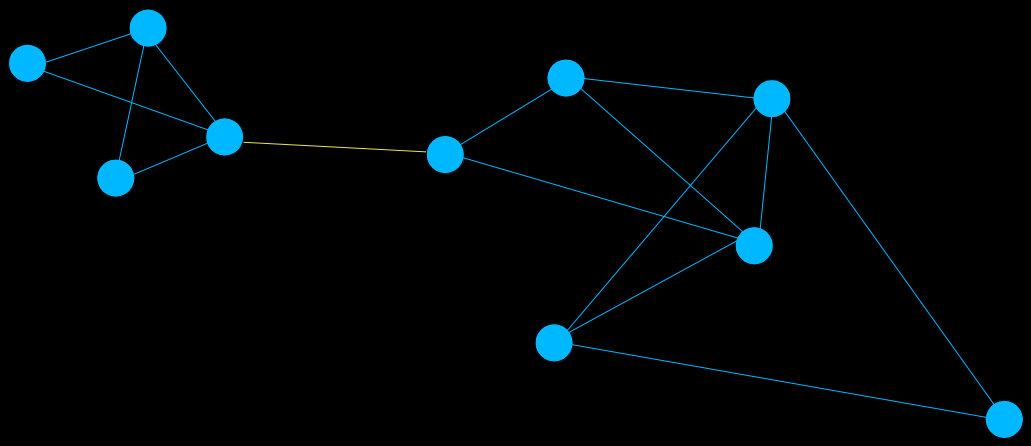
Two dealing with **email data:**

- Tyler, Wilkinson, & Huberman 2003
- Eckmann, Moses, & Sergi 2003

Betweenness and email: Defining betweenness

A measure of centrality, from social network analysis.

$$\text{betweenness}(e) = \sum_{s \neq t} \frac{\#\text{shortest paths}(s,t) \text{ passing through } e}{\#\text{shortest paths}(s,t)}$$

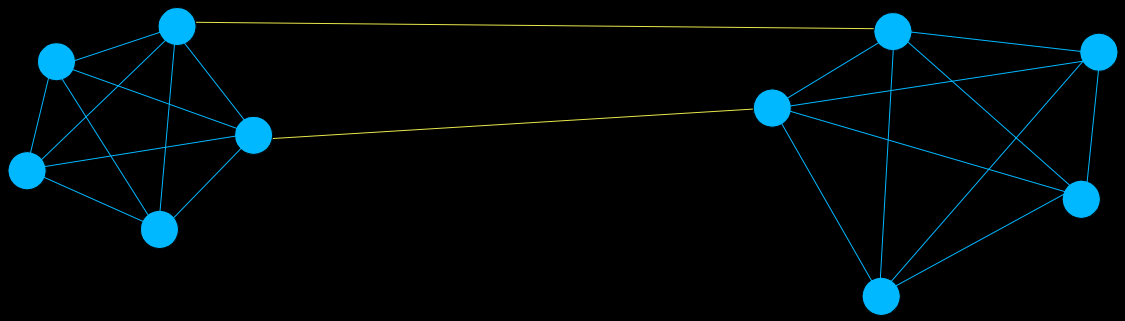


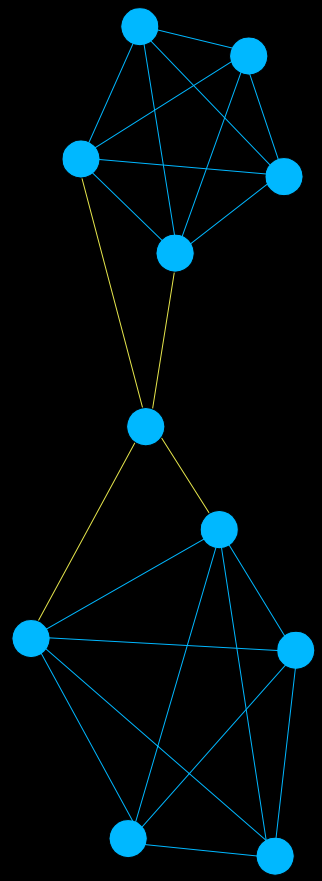
Betweenness and email: Clustering communities

Create an edge when there are
- 30 emails total between parties, **and**
- 5 emails in each direction

while G=(V,E) is nonempty
remove $\text{argmax}_{e \in E} \text{betweenness}(e)$ from E
recalculate betweenness for all edges
for each affected component $C=(V_c,E_c)$
if ($V_c<6$) or ($\max_{e \in Ec} \text{betweenness}(e)<(V_c-1)$)
output C and remove C from G

Betweenness and email: Arbitrary decisions, randomness





The choice of which edge to remove can be arbitrary.

Tyler et al. approximate the maximum betweenness computation, randomly, many times, then aggregate results.

Betweenness and email: Results on the HP Labs graph

485 employees → 367 vertices, 1110 edges

66 communities, from size 2 to size 57
mean: 8.4 members stdev: 5.3

17 crossed organizational boundaries. Formal leaders were near the center.

16 subjects verified the 7 communities they were interviewed about, 7 said people were missing, 6 said an extra was included.

Temporal coupling in the email graph

Three dealing with the **World Wide Web:**

- Gibson, Kleinberg, & Raghavan 1998
- Flake, Lawrence, & Giles 2000
- Adamic & Adar 2003

Two dealing with **email data:**

- Tyler, Wilkinson, & Huberman 2003
- Eckmann, Moses, & Sergi 2003

Temporal coupling: Static graph and curvature

$$\text{Curvature}(v) = \frac{\#\text{edges in neighborhood of v}}{\text{maximum \# of edges neighborhood of v could hold}}$$

Static email graph retains only vertices with curvature greater than 0.1.

Temporal coupling: Mutual information for pairs

$p_A(1)$ = prob. A sends to B = 1-$p_A(0)$
$p_{AB}(0,0)$=prob. neither sends email
$p_{AB}(1,0)$=prob. A sends to B, not vice-versa
$p_{AB}(0,1)$=prob. B sends to A, not vice-versa
$p_{AB}(1,1)$=prob. both send email

Mutual information between a pair of vertices I(A,B)=

$$\Sigma_{i,j=0,1} p_{AB}(i,j) \cdot \log(p_{AB}(i,j) / p_A(i) \cdot p_B(j))$$

Temporal coupling: Mutual information for triads

$$I(A,B,C) = \sum_{i1 \ldots i6 = 0,1} p_{ABC}(i1,i2,i3,i4,i5,i6) \cdot \log\left(\frac{p_{ABC}(i1,i2,i3,i4,i5,i6)}{p_{AB}(i1,i2) \cdot p_{AC}(i3,i4) \cdot p_{BC}(i5,i6)} \right)$$

Temporal coupling: The conjugate graph and time

Create a new graph G’ where triangles in G with temporal cohesion ≥ 0.5 are vertices.

If the corresponding triangles in G share an edge, add an edge in G’.

The result had new clusters, not in the original graph, often across department boundaries.

Common problems with community discovery so far

"Evaluation by admiration" rather than checking against social realities.

Free parameters that cover unknowns about communication.

Difficultly in comparing structural and textual results.

Important advances in community discovery so far

Clear case for the importance of information and communications structure.

Mathematical advances that “boost” understanding of the Web/email/community.

Ethnographic/SNA connections between analysis of data and people’s understanding of their communities.