Geography and Friendship

Joint work with:
David Liben-Nowell: Carleton College
Ravi Kumar, Jasmine Novak, Prabhakar Raghavan: Yahoo! Research
Daniel Gruhl: IBM
Ramanathan Guha: Google

Work performed at IBM, Verity, Yahoo!, Carleton
Some social networks in Yahoo!

- MyWeb 2.0
  - Friendship network
- Instant messenger
  - Buddy list
- Flickr
  - Photo sharing and tagging
- Yahoo!
  - Topically focused communities
What can be studied?

- Structural analysis
- Understanding social phenomena
- Information propagation and diffusion
- Prediction (buzz, information, social)
- Modeling
A study of blogs

- Joint work with:
  - Dan Gruhl (IBM)
  - R. Guha (Google)
  - Ravi Kumar (Yahoo!)
  - David Liben-Nowell (Carleton)
  - Jasmine Novak (Yahoo!)
  - Prabhakar Raghavan (Yahoo!)

- WWW May 2003; CACM Dec 2004; PNAS Aug 2005; KDD Aug 2005; WIP
Etymology

From the OED new ed. (draft entry, Mar 2003) ...

**blog** *intr.* To write or maintain a weblog. Also: to read or browse through weblogs, esp. habitually.

**weblog** *n.* 2. A frequently updated web site consisting of personal observations, excerpts from other sources, etc., typically run by a single person, and usually with hyperlinks to other sites; an online journal or diary.

**blogspace** *n.* The collection of weblogs; = blogosphere, blogsphere, blogistan, …
Blogs 101

- **Characteristics**
  - Pages with reverse chronological sequences of dated entries
  - Usually contain a persistent sidebar containing profile (and other blogs read by the author – “blogroll”)
  - Usually maintained and published by one of the common variants of public-domain blog software

- **From Slashdot, 1999**
  “… a new, personal, and determinedly non-hostile evolution of the electric community. They are also the freshest example of how people use the Net to make their own, radically different new media”
Look and feel

- Quirky
- Highly personal
- Consumed by a small number of regular repeat visitors
- Often updated multiple times each day
- Highly interwoven into a network of small but active micro-communities
- Eg: LiveJournal, Blogger, …
The blog era

- Blogs began in 1996, but exploded in popularity in 1999
  - Proliferation of authoring tools
- Newsweek 2002 estimates ~500K
- Annual Blogathon for charity
  - Bloggers update their Blogs every 30m for 24h
  - Sponsors pay …
- Impact of blogs
  - “Miserable failure”, “French military victories”
Livejournal blogspace

- Livejournal.com: popular blog site
- 1.3M bloggers (Feb 2004)
- 3.9M bloggers (Oct 2005)
- Each blogger has a profile
  - Name, age, …
  - Geographic information (city, state, zip, …)
  - Friends and friend of
  - Interests/communities
Eg, LiveJournal user “bill”

User: **bill** (3215)

Name: bill

Website: Girvan Attractions on the Net

Location: Girvan, United Kingdom

Birthdate: 1954-04-12

E-mail: b.caddis@btinternet.com

Friends: 3: ajose, webfran, zaxwrit

Friend of: 36: agdale, ajose, b4_darkness, boris_the_blade, dkm977, epitaph87, farthead, flatland83, gabbymoe, ghettofabulous, glenda, glitzysgurl, goooooooooooolge, gothgrouch, gruntbill, hammerman, insanephycopath, jakup, jazzzman, laxprincess, louwleadvocals, mandaj8705, marksantos, mini_skeeby, protogonoi, reallyrandom06, sammeh, shortstac, sweetsugar829, sys_developer, thebluesbros, uglyo, uno_bitch, webfran, wikitmel, xo_krista_ox

Member of: 1: paidmembers

**Account type:** Early Adopter

*(more details...)*
LJ bloggers in US
LJ bloggers world-wide
<table>
<thead>
<tr>
<th>Age</th>
<th>%</th>
<th>Representative interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 3</td>
<td>0.5</td>
<td>treats, catnips, daddy, mommy, purring, mice, playing, napping, scratching, milk</td>
</tr>
<tr>
<td>13 to 15</td>
<td>3.5</td>
<td>webdesigning, Jeremy Sumpter, Chris Wilson, Emma Watson, T. V., Tom Felton, FUSE, Adam Carson, Guyz, Pac Sun, mall, going online</td>
</tr>
<tr>
<td>16 to 18</td>
<td>25.2</td>
<td>198{6,7,8}, class of 200{4,5}, dream street, drama club, band trips, 16, Brave New Girl, drum major, talkin on the phone, highschool, JROTC</td>
</tr>
<tr>
<td>19 to 21</td>
<td>32.8</td>
<td>198{3,5}, class of 2003, dorm life, frat parties, college life, my tattoo, pre-med</td>
</tr>
<tr>
<td>22 to 24</td>
<td>18.7</td>
<td>198{1,2}, Dumbledore’s army, Midori sours, Long island iced tea, Liquid Television, bar hopping, disco house, Sam Adams, fraternity, He-Man, She-Ra</td>
</tr>
<tr>
<td>25 to 27</td>
<td>8.4</td>
<td>1979, Catherine Wheel, dive bars, grad school, preacher, Garth Ennis, good beer, public radio</td>
</tr>
<tr>
<td>28 to 30</td>
<td>4.4</td>
<td>Hal Hartley, geocaching, Camarilla, Amtgard, Tivo, Concrete Blonde, motherhood, SQL, TRON</td>
</tr>
<tr>
<td>31 to 33</td>
<td>2.4</td>
<td>my kids, parenting, my daughter, my wife, Bloom County, Doctor Who, geocaching, the prisoner, good eats, herbalism</td>
</tr>
<tr>
<td>34 to 36</td>
<td>1.5</td>
<td>Cross Stitch, Thelema, Tivo, parenting, cubs, role-playing games, bicycling, shamanism, Burning Man</td>
</tr>
<tr>
<td>37 to 45</td>
<td>1.6</td>
<td>SCA, Babylon 5, pagan, gardening, Star Trek, Hogwarts, Macintosh, Kate Bush, Zen, tarot</td>
</tr>
<tr>
<td>46 to 57</td>
<td>0.5</td>
<td>science fiction, wine, walking, travel, cooking, politics, history, poetry, jazz, writing, reading, hiking</td>
</tr>
<tr>
<td>&gt; 57</td>
<td>0.2</td>
<td>death, cheese, photography, cats, poetry</td>
</tr>
</tbody>
</table>
Friendship graph

- Directed
- 80% mutual
- Average degree ~ 14
- Power law degrees
- Clustering coeff. ~ 0.2
- Most friendships explained by age, location, interest
Blogs as trend indicators

- Can blogs be used to predict trends?
- Data
  - Amazon sales rank of some books
  - Blog chatter in an index
- Questions
  - How well do they correlate?
  - Can sales rank be predicted using blogs automatically?
The Lance Armstrong Performance Program
Cross-correlation for Lance Armstrong
Simple inferences

- How to formulate queries automatically
  - Depends on the object (book, CD, DVD, …)
    - Simple heuristics work well
- Predicting sales motion is hard
- Predicting spikes appears relatively easier

- More to be done …
Another question:

- How does friendship depend on geographic distance?
Dataset

- 1.3M LiveJournal bloggers, as of February 2004
- 500K list a home town in the United States
- Home towns mapped to lat/long
- Granularity of locations: roughly cities
- Extracted self-reported “friends” of each blogger: 4M friendships
- 80% of friendships are reciprocal
- ¾ of network form giant strongly-connected component
- Clustering coefficient: 0.2
- Lognormal degree distribution
- Each blogger has a profile
  - Name, age, …
  - Geographic information (city, state, zip, …)
  - Friends and friend of
  - Interests/communities
What’s surprising about Milgram?

- Surprising fact number one (observed by Milgram): network contains short paths
- Surprising fact number two (observed much later by Kleinberg): a purely local algorithm allows discovery of these short paths
Models to explain greedy routing

- Each grid point is a person
- Each person “knows” the four neighbors
- Each person also knows one other person

[Kleinberg 2000]
How should the “long-range” neighbor be chosen

- For a candidate neighbor $x$ at distance $d$ away,
  $\Pr[x \text{ is the long-range neighbor}] \sim 1/d^k$
- If $k=2$:
  - Network contains short paths for every pair (polylog(n))
  - Short paths can be discovered by local greedy routing
- If $k \neq 2$:
  - Networks does not contain short paths (poly(n))
- Exponential gap between $k=2$ and $k\neq 2$
Simulating geographic greedy routing on LiveJournal data

- Can simulate geographic greedy routing on the LiveJournal network
- Results show short paths between most pairs – similar to Milgram’s experiment
- So relationship between friendship and distance should follow $1/d^2$
Results

\[ \varepsilon + 1/d \]

\[ 1/d \]

Kleinberg \((1/d^2)\)
What’s happening?

- Assumption: one person per grid point
- Reality: highly varying number of people per grid point
Population density

- Dot for every inhabited location
- Each circle represents 50,000 bloggers
- Centered on Ithaca, NY
Does population density (or other factors) impact the relationship between friendship and geography?
Our solution

- Why use distance to determine friendship probabilities?
  - Two people who live a mile apart in Beijing will never meet
  - Two people who live a mile apart in Iowa will be close acquaintances
- What’s the difference?
  - Within Manhattan, there are thousands of people living within a mile
  - Within Iowa, there are very few
- Probability of friendship should depend on the size of the candidate population

Pr[friendship] \sim \frac{1}{(# \text{ of closer people})}
Properties of Rank-based friendship

- Population density determines relationship between distance and friendship

- For uniform density, rank-based friendship is equivalent to Kleinberg – same theorems hold
- For non-uniform density, a similar theorem can be shown…
Theorem

- For any $n$-person population network, for arbitrary source $s$, and uniformly-chosen target $t$, the expected length of a geographic greedy routing path from $s$ to the location of $t$ is $O(\log^3 n)$

- Compared to Kleinberg:
  - Lose: expectation rather than with high probability
  - Lose: another log factor
  - Gain: arbitrary population distributions
Generalization 1: General metric spaces

- **Motivation:** “distance” between people may represent complex phenomena: shared interests, similar backgrounds, personality similarity, etc. Would like to allow as general a distance function as possible.

- **Model:**
  - Local edges: pick a shortest path graph in the metric space, include all “local” neighbors that are on a shortest path
  - Long-range edges: rank-based friendship

- **Input:** an n-person social network whose underlying metric space has doubling dimension \( \alpha \), aspect ratio \( AR \), and long-range degree \( d \)

- **Theorem:** For arbitrary source person \( s \) and uniformly chosen target person \( t \), the expected length of a path from \( s \) to the location of \( t \) is \( O(\log(n) \log^2(AR) 2^{\alpha/d}) \).
Generalization 2: Recursive networks

- Motivation: send a message to Manhattan, then route within the sub-network to the correct building, then to the correct room
- Model: As in a standard population network, but each point contains either a singleton person or a recursive sub-network
- Input: a recursive population network of depth $O(poly(n))$
- Theorem: For arbitrary source person $s$ and uniformly chosen destination person $t$, the expected path length from $s$ to $t$ is $O(T \times \min\{\log(n), \text{depth}\})$ where $T$ is the expected path length of a non-recursive network
Generalization 3: Trees with no local edges

- Motivation: many models for social networks have been proposed for trees, without strong routing results
- Input: binary tree of depth $\log^k(n)$
- Model:
  - Each person has $\log^{k+1}(n)$ long-range links by rank-based friendship
  - Local links: none
- Theorem: With arbitrary probability, for arbitrary source person $s$ and uniformly chosen destination person $t$, the expected path length from $s$ to the location of $t$ is $O(\log^k(n))$
Friendship versus rank
East versus West Coast revisited

The diagram shows a plot of link probability against rank, with data points for both West Coast (red triangles) and East Coast (blue circles). The y-axis represents link probability on a logarithmic scale ranging from $1e^{-10}$ to $1e^{-7}$, and the x-axis represents rank also on a logarithmic scale ranging from 100 to 100,000.
How much does geography explain?

- Graph of distance versus friendship probability
- Good estimator of friendship: function of distance plus constant
- Constant term represents geographically-independent reasons for friendship
- Back-solving, we find that 2.5/8 friends are non-geographic
- Could shared interests explain these friendships?
Switching gears: Visualization of Social Networks using Connection Subgraphs

Joint work with:
Christos Faloutsos, CMU
Kevin McCurley, Google

Work performed at IBM Almaden Research Center
Appeared at KDD 2004
Outline

- Introduction / Motivation
- Survey
- Proposed Method
- Algorithms
- Experiments
- Conclusions
Informal Problem Statement

- Given a large social network and two distinguished vertices s and t, show the “relationship” between s and t in the network
- Example: show the relationship between “Nicole Kidman” and “Cameron Diaz”
Standard Approaches

- Standard approach number 1: show an edge if one exists:
  
  Nicole Kidman \rightarrow\llarrow\relbar\relbar\relbar\rightarrow Cameron Diaz
  
  Acted in a movie together

- Standard approach number 2: if no edge exists, show a path:
  
  Nicole Kidman \rightarrow Carmen Electra \rightarrow Cameron Diaz
Proposed Approach

- Show a small subgraph that may capture exponentially many paths concisely:
How big a subgraph?

Given a graph with *initial* and *final* vertices $s$ and $t$, and a budget $B$, return a $B$-node subgraph that best connects $s$ and $t$. 
Budget: 3 nodes
Budget: 5 nodes
Budget: 6 nodes
A larger example: Jan Pedersen to Andrew Tomkins
An example: Byron Dom to David Filo
Fragment of Gary Flake to Bill Gates
Problem definition

- Given a graph, and two nodes $s$ and $t$, and a 'budget' $b$ of nodes
- Find the best $b$ nodes that capture the relationship between $s$ and $t$
Problem definition

- Given a graph, and two nodes $s$ and $t$, and a 'budget' $b$ of nodes
- Find the best $b$ nodes that capture the relationship between $s$ and $t$
Problem definition

- Part 1: How to quantify the goodness?
- Part 2: How to pick ‘best few’ nodes?
- Part 3: Scalability: large graphs (10**7 nodes)
Survey

- Graph Partitioning
  - [Karypis+Kumar]; [Newman+];
  - etc
- Communities
  - [Flake+]; [Kumar, Kleinberg+]
- External distances [Palmer+]
Outline

- Introduction / Motivation
- Survey
- Proposed Method
- Algorithms
- Experiments
- Conclusions
Proposed method for selecting a subgraph

- part 1: measuring quality of a path:
  - electrical current / random walks
- part 2: selecting a subgraph
  - dynamic programming
- part 3: scalability
  - heuristics
Path quality, part 1

- Why not shortest path?
Path quality, part 2

- Why not shortest path?
- Why not net. flow?
Path quality, part 3

- Why not shortest path?
- Why not net. flow?
- Why not plain ‘voltages’?
Path quality, part 4

- Why not shortest path?
- Why not net. flow?
- Why not plain ‘voltages’?
Proposed path quality measure

- Proposed method: voltages with universal sink:
  - ~ ‘tax collector’
- goodness of a path:
- its electric current(*)!
Outline

- Introduction / Motivation
- Survey
- Proposed Method
- Algorithms
- Experiments
- Conclusions
Electricity – Algorithm

- Voltages/Amperages can be computed easily (O(E))
- without universal sink:

\[ v(i) = \sum_j [v(j) \times \frac{C(i,j)}{C(i,*)}] \]

\( i \neq \text{source, sink} \)

\( v(\text{source}) = 1; \ v(\text{sink}) = 0 \)
Electricity – Algorithm

With universal sink:

\[ \nu(i) = \frac{1}{1+a} \sum_j [\nu(j) \times \frac{C(i,j)}{C(i,*)}] \]

(\sim insensitive to \(a = 1\))
Part 2: From paths to subgraphs

- Using Part 1, compute an s-t flow on the entire graph
- Find a subgraph that “captures” much of this flow

Given the flow above, how good is the specified path?
- “Delivered current”: how many electrons travel from s to t along that path
Delivered current of a subgraph

- All units of flow (i.e., electrons) that travel from s to t via edges in the subgraph:
Algorithm for selecting subgraph

- Combinatorial problem: find a B-node subgraph to optimize delivered current – hard to solve exactly or provide approximation algorithms
- Dynamic program to compute:
  - Path which maximizes delivered current per node
- Recursive greedy application
Given the voltages and currents
- Which $b$ nodes to keep?
Part 2: DisplayGen

- ‘delivered current’ of a path:
  - ~ ‘how many electrons’ choose this path

- Diagrams of possible paths and their 'delivered currents'.
Part 2: DisplayGen
Part 2: DisplayGen

- find path to maximize marginal delivered current per node
  - Dynamic programming
- Incrementally, add paths to solution
Part 3: Scalability

Begin with enormous out-of-core graph
Slowly expand from s and t to find a candidate subgraph for algorithm:

Begin with nodes s and t in expansion pool
Until (stoppingCriterion)
   Use pickHeuristic() to pick a node n from expansion pool
   Add n to candidate subgraph
   Add neighbors of n to expansion pool
Apply electrical flow and dynamic program to candidate subgraph
Part 3: Scalability

- By successive, careful expansions
Part 3: Scalability
Part 3: Scalability
Part 3: Scalability
Pseudo-code

Until \texttt{(stoppingCriterion)}
  use \texttt{pickHeuristic()} to pick a node \emph{n}
  expand node \emph{n}
Pseudo-code

`pickHeuristic()` favors
- Nearby nodes with
  - Strong connections to source or sink
  - Small degree
Outline

- Introduction / Motivation
- Survey
- Proposed Method
- Algorithms
- Experiments
- Conclusions
Experiments

- on large real graph
  - ~15M nodes, ~100M edges, weighted
  - ‘who co-appears with whom’ (from 500M web pages)
- Q1: Quality of ‘voltage’ approach?
- Q2: Speed/accuracy trade-off?
Q1: Quality

- Actors (A); Computer-Scientists (CS)
- Kidman-Diaz (A-A)
- Negreponte-Palmisano (CS-CS)
- Turing-Stone (CS-A)
What are the best paths between ‘Kidman’ and ‘Diaz’?
• Mainly: CEOs of major Computer companies (Dell, Gates, Fiorina, +++)
CS-CS: Negroponte - Palmisano

NN  Esther Dyson  Louis Gerstner  SP
CS-A: Turing - Stone
Outline

- Introduction / Motivation
- ...
- Experiments
  - Q1: quality
  - Q2: speed/accuracy trade-off
- Conclusions
Speed/Accuracy Trade-off

delivered current

number of nodes kept (‘b’)

Kleinberg-Newell
Rivest-Hoffman
Turing-Stone
Kidman-Diaz
Speed/accuracy trade-off

- 80/20-like rule:
- the first few nodes/paths contribute the vast majority of ‘delivered current’
- Thus: CandidateGen makes sense
Conclusions

- Defined the problem
- Part 1: Electricity-based method to measure quality
- Part 2: Dynamic programming to spot best paths (‘DisplayGen’)
- Part 3: Scalability with good accuracy (‘CandidateGen’)
- Operational system
Conclusions

- Friendship and Distance are strongly related
- Modeling friendship as a function of distance is problematic
- Rank is a better measure of friendship than distance
- Some friendships form with no geographic correlation (2.5/8)
More Information

- Email: atomkins@yahoo-inc.com
- Web: http://www.tomkinshome.com/andrew