Two Fun Topics + Quick Overview

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu
Users & Online Communities

Online Communities:

- Understanding the co-evolution of users and communities
- Steering user behavior

12/3/2014
Online Communities

Modeling the relation between a user and a community
Users and Communities

Questions:

- How does a user become member of a community?
- How do user and community practices co-evolve?
- Can we predict when a user will leave the community?
Language practices (norms, etiquette, ...)  
- build collective identity  
- foster individual expression  

Linguistic change captures the relation between users and communities  
- Framework for tracking linguistic change  
- Measures of user reaction to linguistic change  
- Predict when user will leave the community
Online Communities

- **People discuss beer:**

  ![Example screenshot from RateBeer](image)

  Clear copper colored brew, medium cream colored head. **Floral** hop nose, caramel malt. Caramel malt front dominated by a nice **floral** hop background. Grapefruit tones. Very tasty hops run the show with this brew. Thin to medium mouth. Not a bad choice if you’re looking for a nice hop treat.

- **10 years of complete linguistic data**
  - **RateBeer:** 3M posts, 30k users
  - **BeerAdvocate:** 1.6M posts, 33k users
“Aroma” vs. “Smell”

The graph shows the usage of the terms 'Aroma' and 'Smell' over a period from 2001 to 2011. "Aroma" convention peak in 2003, while "Smell" convention usage remains relatively consistent from 2005 onwards.
Users who joined in 2003

New users are more likely to use "smell" than users who have been part of the community for a long time.
Community / User Change

User:

“life stage”

Community:
(a) First person sing. pronouns    (b) Beer specific vocabulary

- A sign of increasing identification with the community [Pennebaker 2007; Sherblom 2009]
Fruit words (*peach, pineapple, berry, …*) are getting ever more popular
**User-Community Change**

Language distance between the user and the community:

- **Snapshot Language Model** ($SLM_t$): Lang. model of the community at time $t$
- **Distance**: cross-entropy of $U_t$ according to $SLM_t$:

$$H(U_t, SLM_t) = -\frac{1}{N} \sum_i \log P_{SLM_t}(w_{u_t,i})$$
Stage 1: User assimilates language of the community
User-Community Change

Time

Community language

Distance from the community

Life stage

5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55% 60% 65% 70% 75% 80% 85% 90% 95% 100%

User

Stage 2: User’s language distances itself from that of the community
User-Community Change

**H1:** User moves away

**H2:** User stops adapting

![Graph showing community language change over time with distance from the community on the y-axis and life stage on the x-axis.](image-url)
H2: Users get stuck in the past!

Self-similarity of user’s language at time $t$ and $t-1$
Elastic Lifecycle

- So far we stretched lifetimes to 0-100%
- What about user’s absolute lifetime?

→ Similar lifecycle in spite of different lifespans
   “All users die old”
→ End of the adaptation phase is a function of the ultimate lifespan
→ Level of receptivity is related to the ultimate lifespan
Elastic Lifecycle

- So far we stretched lifetimes to 0-100%
- What about user’s absolute lifetime?

→ Similar lifecycle in spite of different lifespans
  “All users die old”
→ End of the adaptation phase is a function of the ultimate lifespan
→ Level of receptivity is related to the ultimate lifespan

Predict user’s ultimate lifespan
How do we influence social systems?
People work amazingly hard to earn badges
“Give me enough medals and I’ll win you any war.”

– Napoleon
**BADGE TYPES**

**Meteorite** badges are common and easy to earn when just getting started.

**Moon** badges are uncommon and represent an investment in learning.

**Earth** badges are rare. They require a significant amount of learning.

**Sun** badges are epic. Earning them is a true challenge, and they require impressive dedication.

**Black Hole** badges are legendary and unknown. They are the most unique Khan Academy awards.

**Challenge Patches** are special awards for completing topic challenges.
<table>
<thead>
<tr>
<th>Role</th>
<th>Reputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author</td>
<td>2694</td>
</tr>
<tr>
<td>Analytical</td>
<td>421</td>
</tr>
<tr>
<td>Announcer</td>
<td>13</td>
</tr>
<tr>
<td>Archaeologist</td>
<td></td>
</tr>
<tr>
<td>Autobiographer</td>
<td></td>
</tr>
<tr>
<td>Benefactor</td>
<td>14</td>
</tr>
<tr>
<td>Beta</td>
<td>2525</td>
</tr>
<tr>
<td>Booster</td>
<td>744</td>
</tr>
<tr>
<td>Caucus</td>
<td>85162</td>
</tr>
<tr>
<td>Citizen Patrol</td>
<td></td>
</tr>
<tr>
<td>Civic Duty</td>
<td>21</td>
</tr>
<tr>
<td>Cleanup</td>
<td>1300</td>
</tr>
<tr>
<td>Commentator</td>
<td></td>
</tr>
<tr>
<td>Constituent</td>
<td>21</td>
</tr>
<tr>
<td>Convention</td>
<td>1</td>
</tr>
<tr>
<td>Copy Editor</td>
<td>7</td>
</tr>
<tr>
<td>Critic</td>
<td>106518</td>
</tr>
<tr>
<td>Custodian</td>
<td>596</td>
</tr>
<tr>
<td>Deputy</td>
<td>2322</td>
</tr>
<tr>
<td>Editor</td>
<td>473573</td>
</tr>
<tr>
<td>Electorate</td>
<td>3426</td>
</tr>
<tr>
<td>Enlightened</td>
<td>800</td>
</tr>
<tr>
<td>Enthusiast</td>
<td>4533</td>
</tr>
<tr>
<td>Epic</td>
<td>352</td>
</tr>
<tr>
<td>Excavator</td>
<td>2911</td>
</tr>
<tr>
<td>Famous Question</td>
<td></td>
</tr>
<tr>
<td>Fanatic</td>
<td>7992</td>
</tr>
<tr>
<td>Favorite Question</td>
<td></td>
</tr>
<tr>
<td>Generalist</td>
<td>466</td>
</tr>
<tr>
<td>Good Answer</td>
<td>51</td>
</tr>
<tr>
<td>Good Question</td>
<td></td>
</tr>
<tr>
<td>Great Answer</td>
<td>6</td>
</tr>
<tr>
<td>Great Question</td>
<td></td>
</tr>
<tr>
<td>Guru</td>
<td>15701</td>
</tr>
<tr>
<td>Informed</td>
<td>21165</td>
</tr>
<tr>
<td>Investor</td>
<td>5386</td>
</tr>
<tr>
<td>Marshal</td>
<td>480</td>
</tr>
<tr>
<td>Mortarboard</td>
<td>13183</td>
</tr>
<tr>
<td>Necromancer</td>
<td>7193</td>
</tr>
<tr>
<td>Nice Answer</td>
<td>2676</td>
</tr>
<tr>
<td>Nice Question</td>
<td>113</td>
</tr>
<tr>
<td>Notable Question</td>
<td></td>
</tr>
<tr>
<td>Organizer</td>
<td>45228</td>
</tr>
<tr>
<td>Outspoken</td>
<td>411</td>
</tr>
<tr>
<td>Peer Pressure</td>
<td>213</td>
</tr>
<tr>
<td>Popular Question</td>
<td></td>
</tr>
<tr>
<td>Populist</td>
<td>2976</td>
</tr>
<tr>
<td>Precognitive</td>
<td></td>
</tr>
<tr>
<td>Promoter</td>
<td>25419</td>
</tr>
<tr>
<td>Proofreader</td>
<td>2862</td>
</tr>
<tr>
<td>Publicist</td>
<td>302</td>
</tr>
<tr>
<td>Pundit</td>
<td>3026</td>
</tr>
<tr>
<td>Quorum</td>
<td>12075</td>
</tr>
<tr>
<td>Research Assistant</td>
<td></td>
</tr>
<tr>
<td>Reviewer</td>
<td>3847</td>
</tr>
<tr>
<td>Revival</td>
<td>80128</td>
</tr>
<tr>
<td>Scholar</td>
<td>45361</td>
</tr>
<tr>
<td>Self-Learner</td>
<td>303</td>
</tr>
<tr>
<td>Sportsmanship</td>
<td></td>
</tr>
<tr>
<td>Stellar Question</td>
<td></td>
</tr>
<tr>
<td>Steward</td>
<td>948</td>
</tr>
<tr>
<td>Stunk &amp; White</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>60236</td>
</tr>
<tr>
<td>Suffrage</td>
<td>14365</td>
</tr>
<tr>
<td>Supporter</td>
<td>3542</td>
</tr>
<tr>
<td>Synonymizer</td>
<td>4</td>
</tr>
<tr>
<td>Tag Editor</td>
<td>8944</td>
</tr>
<tr>
<td>Talkatire</td>
<td>3169</td>
</tr>
<tr>
<td>Taxonomist</td>
<td>610</td>
</tr>
<tr>
<td>Teacher</td>
<td>44264</td>
</tr>
<tr>
<td>Tenacious</td>
<td>138</td>
</tr>
<tr>
<td>Tumbleweed</td>
<td>270</td>
</tr>
<tr>
<td>Unsung Hero</td>
<td>4</td>
</tr>
<tr>
<td>Vox Populi</td>
<td>661</td>
</tr>
</tbody>
</table>
Multiple roles of badges:
- Can recognize a wide range of activities:
  - Total effort, Single high-impact contribution, ...
- Serve both as credentials and incentives

Incentive aspects of badges:
- Trend toward gamification [Deterding et al. ‘11]
- Customer loyalty programs [Kopalle-Neslin ‘03]
- Simple for users to understand, and less based on competition with others
Badges & Behavior Change

How do badges translate into effects on user behavior?

Need lots of data to tease out the effects and build a mathematical model

Badges on Stack Overflow Q&A site:
2M people
5M questions
10M votes

How to format a JSON date?

I'm taking my first crack at Ajax with jQuery. I'm getting my data onto my page, but I'm trouble with the JSON data that is returned for Date data types. Basically, I'm getting looks like this:

```
/Date(1224843280000)/
```

From someone totally new to JSON - How do I format this to a short date format? She handled somewhere in the jQuery code? I've tried the `jQuery UI datepicker formatDate()` without any success.

FYI: Here's the solution I came up with using a combination of the answers here:

This solution got my object from the callback method and display the dates e date format library.

```
/Date(1224843280000)/
```

434 views
48 likes

- [jquery](https://www.jquery.com)
- [asp.net](https://www.asp.net)
- [asp.net-mvc](https://www.asp.net/mvc)
- [ajax](https://www.ajax.com)
- [json](https://www.json.com)

[link](https://example.com) | [edit](https://example.com) | [flag](https://example.com)

*Peter Mortensen*
4,761 views 12/3/12 10:05 PM

*May 5 <1 at 16:09*

Badges on StackOverflow

I am trying to get the list of connected components in a graph with 100 million nodes. For smaller graphs, I usually use the `connected_components` function of the Networkx module in Python which does exactly that. However, loading a graph with 100 million nodes (and their edges) into memory with this module would require ca. 110GB of memory, which I don't have. An alternative would be to use a graph database which has a connected components function but I haven't found any in Python. It would seem that Dex (API: Java, .NET, C++) has this functionality but I'm not 100% sure. Ideally I'm looking for a solution in Python. Many thanks.

**Newbie**: Congrats on your 1st answer

**Superstar**: You answered 10 questions
Model of Badges

- **Approach:** Utility based model of badges
  - User trades off between her preferred mix of activities and the goal of reaching the badge

- **Effects on both engagement and steering**
  - **Engagement:** Increased user site activity
  - **Steering:** Users change the actions they do
2 parts to the model:

- User gains value from obtaining a badge
- But it “costs” user to change behavior
User’s optimization problem:
Choose actions to maximize utility
- User balances between achieving badges quickly and keeping cost low

At each point $x_a$ in time:
- Receive utility $V_b$ of badges already won
- Get penalized for deviating from preferred behavior $p$ according to loss function $g(p, x)$
- Exit the site with probability $\theta$
Model of Badges

\[ U(x_a) = \sum_{b \in B} l_b(a) V_b + \theta \sum_{i=1}^{n+1} x_a^i \cdot U(x_{a+e_i}) - g(x_a, p) \]

- \( U(x_a) \)... total utility for a user with action distribution \( x_a \)
- \( a \)... user’s actions so far (count of actions)
- \( l_b(a) \)... given user’s actions \( a \), did she get badge \( b \) (0/1)
- \( V_b \)... value of badge \( b \)
- \( \theta \)... prob. of exiting the site
- \( x^i_a \)... user’s prob. of taking action \( i \)
- \( U(x_{a+e_i}) \)... future utility after taking action \( i \)
- \( g(x_a, p) = \sum_i (x^i_a - p_i)^2 \)... cost for deviating from \( p \)
- \( p \)... preferred action distribution
Solving the Model

- Solve $\text{arg max } U(x_a) = \sum_{b \in B} l_b(a) V_b + \theta \sum_{i=1}^{n+1} x_a^i \cdot U(x_a + e_i) - g(x_a, p)$

  - Partition space based on badge boundaries
  - Inductively solve regions in order from large coordinates to small
Example: 2 Badges

- Influencing user behavior:

![Graph showing user behavior with badges](image-url)
Example: Badge at 25 actions of type $A_1$

- User steers in $A_1$ direction as she approaches the badge boundary; then resets after receiving it.
Model vs. Data

- Model predicts qualitative behavior

(Electorate) (Vote on at least 600 questions)
Question: How should you “place” badges to achieve desired effects?

Our model allows for optimizing the badge placement for optimal behavior steering:
Badges on Coursera

**The Reader**
To earn the next badge (Silver), you must read 30 threads from your classmates.

**The Supporter**
To earn the next badge (Silver), you must vote on 15 posts that you find interesting or useful.

**The Contributor**
To earn the next badge (Bronze), you must post 3 replies that your classmates find interesting.

**The Conversation Starter**
To earn the next badge (Bronze), you must start 3 threads that your classmates find interesting.

**Top Posts**
To earn the next badge (Bronze), you must write a post that gets 5 upvotes from your classmates.
Experiment: ML Class

5x more likely to get to 100 votes/reads!
Experiment: ML Class

No qualitative difference in posts/comments
No badges on these actions!
Badges: Further questions

- Many questions, both within this framework and extending it:
  - Where does the value of a badge come from? Internal, social, transactional, ...
  - How does achievement-seeking interact with competition and scarcity?
  - How far can we develop analogies with off-line domains?
Social and Information Network Analysis: Review of Key Concepts
How do we reason about networks?
Reasoning About Networks

How do we reason about networks?

- **Empirical:** Study network data to find organizational principles
- **Mathematical models:** Probabilistic, graph theory
- **Algorithms:** Methods for analyzing graphs

What do we hope to achieve from models of networks?

- Patterns and statistical **properties** of network data
- Design principles and **models**
- **Understand** why networks are organized the way they are (Predict behavior of networked systems)
What do we study in networks?

- **Structure and evolution:**
  - What is the structure of a network?
  - Why and how did it become to have such structure?

- **Processes and dynamics:**
  - Networks provide “skeleton” for spreading of information, behavior, diseases
What We Have Covered

- Network diameter
- Edge clustering
- Scale-free networks
- Strength of weak ties
- Core-periphery structure
- Densification power law
- Shrinking diameters
- Structural Balance
- Status Theory
- Memetracking
- Small-world model
- Erdös-Renyi model
- Preferential attachment
- Network cascades

- Independent cascade model
- Decentralized search
- PageRank
- Hubs and authorities
- Girvan-Newman
- Modularity
- Clique percolation
- Supervised random walks
- Influence maximization
- Outbreak detection
- Linear Influence Model
- Network Inference
- Kronecker Graphs
- Bow-tie structure
## How It All Fits Together

### Properties
- Small diameter, Edge clustering
- Scale-free
- Strength of weak ties, Core-periphery
- Densification power law, Shrinking diameters
- Patterns of signed edge creation
- Information virality, Memetracking

### Models
- Small-world model, Erdős-Renyi model
- Preferential attachment, Copying model
- Kronecker Graphs
- Microscopic model of evolving networks
- Structural balance, Theory of status
- Independent cascade model, Game theoretic model

### Algorithms
- Decentralized search
- PageRank, Hubs and authorities
- Community detection: Girvan-Newman, Modularity
- Link prediction, Supervised random walks
- Models for predicting edge signs
- Influence maximization, Outbreak detection, LIM
Small-World Phenomena

- **Properties:**
  - Six degrees of separation
    - Networks have small diameters
  - Edges in the networks cluster
    - Large clustering coefficient

- **Models:**
  - Erdös-Renyi model
    - Baseline model for networks
  - The Small-World model
    - Small diameter and clustered edges

- **Algorithms:**
  - Decentralized search in networks
    - Kleinberg’s model and algorithm
Scale-Free Networks

- **Properties:**
  - Power-law degrees
    - Degrees are heavily skewed
  - Network resilience
    - Networks are resilient to random attacks

- **Models:**
  - Preferential attachment
    - Rich get richer

- **Algorithms:**
  - Hubs and Authorities
    - Recursive: \( a_i = \sum_{j \rightarrow i} h_j, h_i = \sum_{i \rightarrow j} a_j \)
  - PageRank
    - Recursive formulation, Random jumps
Community Detection

- **Properties:**
  - Strength of weak ties
  - Core-periphery structure

- **Models:**
  - Kronecker graphs model

- **Algorithms:**
  - Spectral Clustering
  - Girvan-Newman (Betweenness centrality)
  - **Modularity:** \(\text{#edges within group} - \text{E[#edges within group]}\)
  - Clique Percolation Method
    - Overlapping communities
Network Evolution

- **Properties:**
  - Densification Power Law
    - \( E(t) \propto N(t)^a \)
  - Shrinking Diameter

- **Models:**
  - Microscopic Network Evolution
    - Exponential life-times, Evolving sleeping times
    - Random-Random edge attachment

- **Algorithms:**
  - Link prediction
**Signed Networks**

- **Properties:**
  - Signed link creation
  - +links are more embedded

- **Models:**
  - Structural Balance
    - Coalition structure of networks
  - Status Theory
    - Global node status ordering

- **Algorithms:**
  - Predicting edge signs
Network Diffusion (1)

- **Properties:**
  - Meme-tracking
    - Blogs trail mass media

- **Models:**
  - Game theoretic model:
    - Payoffs, Competing products
  - Independent Cascade Model
    - Each node infects a neighbor with some probability
Algorithms:

- **Influence Maximization**
  - Set of \( k \) nodes producing largest expected cascade size if activated
  - Submodularity
  - Greedy hill-climbing

- **Outbreak Detection**

- **Network Inference**
  - Infer networks based on information diffusion data
## Map of Superpowers

### Properties
- Small diameter, Edge clustering
- Scale-free
- Strength of weak ties, Core-periphery
- Densification power law, Shrinking diameters
- Patterns of signed edge creation
- Viral Marketing, Blogosphere, Memetracking

### Models
- Small-world model, Erdős-Renyi model
- Preferential attachment, Copying model
- Kronecker Graphs
- Microscopic model of evolving networks
- Structural balance, Theory of status
- Independent cascade model, Game theoretic model

### Algorithms
- Decentralized search
- PageRank, Hubs and authorities
- Community detection: Girvan-Newman, Modularity
- Link prediction, Supervised random walks
- Models for predicting edge signs
- Influence maximization, Outbreak detection, LIM
Applying Your Superpowers
People are connected

Technology is helping us to stay connected

We cannot improve the technology and services without understanding human interactions and networks behind them

Can we recognize fundamental patterns of human behavior from raw digital traces?
Intelligence and fighting (cyber) terrorism
Applying Your Superpowers

- Predicting epidemics: Ebola
Applying Your Superpowers

- Interactions of human diseases
- Drug design
What Next?

- **Project write-ups:**
  - Tue Dec 9 Midnight **(11:59PM)** Pacific Time
    - 1 team member uploads PDF to Gradescope
    - 5-10 pages; See course website for more info

- **Poster session:**
  - Thu Dec 11 from 12- 3pm in Packard Atrium
    - All groups with at least one non-SCPD member must present
    - There should be 1 person at the poster at all times
    - Prepare a 2-minute elevator pitch of your poster
    - There will be food

No late days!
What Next? Seminars

- **Seminars:**
  - **InfoSeminar:** [http://i.stanford.edu/infoseminar](http://i.stanford.edu/infoseminar)
    - Can be taken for credit (CS545, also on SCPD)
    - Great industry/academia speakers on Fridays

- **Conferences / Journals:**
  - **WWW:** ACM World Wide Web Conference
  - **WSDM:** ACM Web search and Data Mining
  - **ICWSM:** AAAI Int. Conf. on Web-blogs & Social Media
  - **KDD:** Conf. on Knowledge Discovery & Data Mining
  - **Journal of Network Science**
  - **Journal of Complex Networks**
What Next? Courses

- **CS246: Mining Massive Datasets (Winter 2014)**
  - Data Mining & Machine Learning for big data
    - (big==does’ fit in memory/single machine), MapReduce

- **CS341: Project in Data Mining (Spring 2014)**
  - Groups do a research project on big data
  - We provide interesting data, projects and access to the Amazon computing infrastructure
  - Nice way to finish up CS224W project & publish it!
Other relevant courses:

- CS276: Information Retrieval and Web Search
- CS229: Machine Learning
- CS245: Database System Principles
- CS347: Transaction Processing and Distributed Databases
- CS448g: Interactive Data Analysis