Political Dynamics in Large Scale online Data Sets: A Study of Content-Oriented User Behavior

Roja Bandari
Big Data, Little Insight

- Growing shift to online interaction is affecting society
  - Example: opinion formation and decision making
  - Government, Industry, Academia have taken notice
- User → Thousands of small actions (tweets, likes, comments, clicks,)

**Accel Launches $100M Big Data Fund 2 To Invest In The ‘Second Wave’ Of Big Data Startups**

"How can we have this much data and still not understand collective human behavior?"
Background

- Explicit (e.g. friendships)
- Implicit (e.g. interactions)

Data → Network model → Detect patterns → Predict → Problems

- Methods/Approaches:
  - Supervised, Unsupervised, Probabilistic, Deterministic, etc.

- Examples:
  - Friendship recommendations
  - Prediction of popularity (e.g. films)
Promise for Social Sciences

- Access to highly granular, time-stamped data
- Datasets raise hopes for data-driven models in social sciences
  - Scale, complexity, and noisiness
  - Predicated on automated methods to extract *summative macroscopic observables*
    - Equivalents of pressure and temperature
- Current approaches:
  - Summarizing the network: e.g. graph-theoretic communities
  - Summarizing the content: e.g. Topic Modeling
    - Hidden topics derived through word co-occurrence across documents
    - Computationally intensive
    - Mathematical regularization rather than social regularization
Focus of Talk

- Content oriented collective behavior
- Macroscopic observables to extract:
  - *Collective behavior*: different meanings across many traditions.
    - Here we use it in a general sense: attitudes and actions of groups of users
  - Collective behavior in social news
- *Content-oriented*: involving users’ creation, sharing, and promotion of content and their attitudes toward content (text or other material)
Social News

- Social news aggregation sites have article submission, voting, and commenting capabilities
  - Examples: Reddit, Digg, Slashdot, Balatarin
Motivating Questions

- How would one begin to understand the user population?
- Theory of structuration  (Anthony Giddens)
  - Emphasizes both agency and structure

If these structures exist they must manifest themselves in aggregated data.

Once we can detect macro structures (collective behavior), we can answer other questions:
  - Do users form polarized and insular groups?
  - Does one group dominate or drive out other groups?
  - How do external events affect these dynamics?
Dataset and Methodology

- **Balatarin**: popular Persian-language social news site
  - 4 years of data: users, articles, and user votes to articles
  - Politics Category: 26,000 users 350,000 articles, 9.2M votes

- **Votes**: user actions
  - Explicit indicators of user preference for content
  - Other examples: *Like* (Facebook), *up-vote* (Reddit), +1 (G+), Digg

- Detect communities of users with similar voting patterns and track these communities' temporal evolution.

- Characterize evolving communities through their preferred content
Methodology

Divide data into periods → Detect user communities in each period → Map consecutive communities → Extract representative articles → Extract representative terms and domains

Divide data into consecutive overlapping time periods (30 days, 14 day overlap).

Path summary:
- Representative Domains
- Representative Words
Communities

- Higher density of edges within communities than between them
- Modularity* = fraction of edges that belong to the same community in the graph minus the null model

\[ Q = \frac{1}{2m} \sum_{x,y} (A_{xy} - \frac{k_x k_y}{2m}) \delta(C^x, C^y) \]

- Null model:
  - Graph with same degree sequence
  - Connect pairs of edge stubs (2m) at random
  - Optimize by iteratively joining communities, starting with single-node communities.

* Developed by Girvan, Clauset, Newman
Bipartite Projection

- In each time period votes create a bipartite graph of articles and users
- Project to a weighted unipartite network

\[ W_{jaccard} = \frac{n(X \cap Y)}{n(X \cup Y)} \]

Weight between users x and y
X: set of articles voted for by user x
Y: set of articles voted for by user y
n: set cardinality
Detect and Map Communities

- For a weighted graph
  - Replace $A_{xy}$ with $W_{xy}$ and $m$ with total weight in the graph, $W$.
  - Replace vertex degree $k_x$ with vertex strength $s_x$

$$Q = \frac{1}{2W} \sum_{x,y} (W_{xy} - \frac{s_x s_y}{2W}) \delta(C^x, C^y)$$

$$s_x = \sum_y W_{xy}$$

- Communities reflect users with similar content preference
- Map consecutive communities based on user overlaps.

Time = $t_1$

$C_i$

0.5

0.04

Time = $t_2$

$C_j$

Users in community $i$

$$\frac{n(C_i \cap C_j)}{n(C_i)} \cdot \frac{n(C_i \cap C_j)}{n(C_j)}$$
Define an Evolution Path

- Define a path as consecutive mapping of communities with no merges or splits lasting a minimum duration (at least 3 months long)
- Size of each oval represents size of community
In each time window, find articles that are highly preferred by each community.

Assuming each community votes for articles at random with probability:

$$p_i = \frac{N_i}{N}$$

Then probability that $o_{ij}$ of an article’s $N_j$ votes come from community $i$:

$$p(o_{ij}) = \binom{N_j}{o_{ij}} p_i^{o_{ij}} (1 - p_i)^{(N_j - o_{ij})}$$

For $o_{ij} > p_i \cdot N_j$, the lower this probability, the higher the preference of community $i$ for article $j$. 
Representative Terms and Domains

- **Representative terms** (in articles preferred by a community)

\[
\text{Score}(T) = \frac{\text{tf}(T, C)}{\max_i \text{tf}(t, C)} - \frac{\text{tf}(T)}{\max_t \text{tf}(t)}
\]

- Aggregate these terms as well as the websites of preferred articles over each path

<table>
<thead>
<tr>
<th>Domains</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.bbc.co.uk">www.bbc.co.uk</a></td>
<td>Minister, Nuclear, Spokesperson, Russia</td>
</tr>
<tr>
<td><a href="http://www.dw-world.de">www.dw-world.de</a></td>
<td>Council, Continuation, Israel, Security</td>
</tr>
<tr>
<td><a href="http://www.roozonline.com">www.roozonline.com</a></td>
<td>Iraq, Arrangement, Agency, Europe</td>
</tr>
<tr>
<td>radiozamaaneh.com</td>
<td>America, Declare</td>
</tr>
<tr>
<td><a href="http://www.radiofarda.com">www.radiofarda.com</a></td>
<td></td>
</tr>
</tbody>
</table>
The Green Movement

Massive protests and their subsequent violent crackdown coincide with a major change in dynamics of the site.

G: Iran’s foreign affairs: nuclear talks, US, Russia,

- A: Reformist
- B: Conservative
- C: Weak Conservative
- D: Anti-Ahmadinejad
- E: Sarcastic Opposition
- F: Foreign Affairs 1
- G: Foreign Affairs 2
- H: Sensationalist
- I: Green Protest 1
- J: Anti-Reformist
- K: Green Protest 2
- L: Green Protest 3
- M: Separatist
- N: Green Human Rights

N: Articles about human rights violations committed against the protesters by the government and the arrests of activists.

B: Articles published by conservative fundamentalist websites.

I, K, L: Large pro-Green Movement consecutive paths. YouTube videos of protests and eyewitness accounts are the focal points.

J: Against prominent reformist figures.
Principal Component Analysis Corroborates Path Meanings

- PCA plot of core-user overlaps
- A temporal and a political dimension emerge as a result of PCA analysis on user overlaps in paths.
- First two components explain 43% of variance
- Contents of paths agree with path positions in the PCA political dimension.
Domain Recurrence

- Are some domains repeatedly preferred in a path?
  - Aggregate domains over whole path and count their recurrence
  - Compare with count of domains if the votes were drawn at random
    - Draw votes at random and note their domains

- Higher relative recurrence = more uniformity in domains

$$\text{Entropy}(C) = - \sum_{i} p_i \log_2(p_i)$$

\(p_i\): Probability that an article from domain \(i\) is in the top \(n\) most preferred articles of a path.

Paths with high recurrence:
B(conservative), G(foreign affairs), K(eyewitness)
User Retention

Retention($P$, $\Delta \tau$) = \frac{n(P(\tau_i) \cap P(\tau_i + \Delta \tau))}{n(P(\tau_i))}

where $P(\tau_i)$ is the set of users in path $P$ at time $\tau_i$.

- Paths with high retention: G, N, K
- Paths with low retention: B, C, E, J, H
Parameter variations

- Three parameters were chosen:
  - $W$: Window length for each time period
  - $S$: Shift length determines overlap between consecutive windows
  - $\text{Th}$: Threshold for elimination of low-vote users
- Overlapping windows of size $W$ shifted $S$ days at each period

- More paths: higher granularity
- Longer path: more consistency, easier interpretation
- Prefer more and longer paths

![Graph showing number of paths and average length of paths for different parameter variations.](image)
Alternative dataset: sports

- Method is applicable to different contexts
- Found that Sports is highly event-driven with some early adopters for each event, joined by the rest some periods later:
  - Asian (soccer) cup
  - National leagues
  - European cup
  - Paralympics
Gestalt Computing

- A macro structure
  - The parts create the whole but the whole adds to the parts \(\rightarrow\) more than the summation of its parts.
  - Constructed from elementary user actions
  - More than sum of its parts:
    - Relationship between parts of the structure.
    - What is not there as well as what is there.

From the Merriam-Webster dictionary: Gestalt is a structure, configuration, or pattern of physical, biological, or psychological phenomena so integrated as to constitute a functional unit with properties not derivable by summation of its parts.

- We began with elementary actions (votes) \(\rightarrow\) obtained global structure \(\rightarrow\) the context in the structure gives back meaning to individual actions.
Structure Reveals a New Perspective

- Comparing two users: 2 of their top 20 domains are different.
  - **User 1 Simpson Index: 0.41**
    - Core users in paths A, F, G, N: Reformist, Foreign affairs, Human rights.
  - **User 2 Simpson Index: 0.34**

<table>
<thead>
<tr>
<th>Domain</th>
<th>User 1 Activity</th>
<th>User 2 Activity</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.youtube.com">www.youtube.com</a></td>
<td>13673</td>
<td><a href="http://www.youtube.com">www.youtube.com</a></td>
<td>1042</td>
</tr>
<tr>
<td><a href="http://www.bbc.co.uk">www.bbc.co.uk</a></td>
<td>9012</td>
<td><a href="http://www.bbc.co.uk">www.bbc.co.uk</a></td>
<td>517</td>
</tr>
<tr>
<td><a href="http://www.radiofarda.com">www.radiofarda.com</a></td>
<td>7264</td>
<td><a href="http://www.radiofarda.com">www.radiofarda.com</a></td>
<td>430</td>
</tr>
</tbody>
</table>

User 1 is more consistent

\[
\text{Simpson's index} = \sum_{i \in A:N} p_i^2
\]

Proportion of a user's activity that is in path \( i \)
In Summary:

- Automated and unsupervised
  - Deriving the structure requires no expert knowledge of the forum under study

- Paths with distinct and meaningful preferences.

- Incorporates both users and content (vs. just one)

- Reveals a new perspective otherwise unknown

- Applicable to other contexts

- **Path width:** number of unique users in the path.
- **Arrows:** inter-path migrations.
- **Darkness:** user retention.
Concluding Remarks

- Automated and unsupervised method produced political paths with distinct and meaningful preferences.

- Questions:
  - Does the approach sacrifice complexity and sophistication?
  - Is this the “single” “True” structure?
  - Can one combine user actions with different/multiple/undefined intentions?

- Ethical considerations: surveillance and privacy
Thank You