Social Media Mining

Behavior Analytics
Examples of Behavior Analytics

• What motivates individuals to join an online group?

• When individuals abandon social media sites, where do they migrate to?

• Can we predict box office revenues for movies from tweets posted by individuals?
• To answer these questions we need to analyze or predict behaviors on social media.

• Individuals exhibit different behaviors in social media: as individuals
• or as part of a broader collective behavior.

• When discussing individual behavior, our focus is on one individual.

• Collective behavior emerges when a population of individuals behave in a similar way with or without coordination or planning.
Our Goal

To **analyze**, **model**, and **predict** individual and collective behavior
Individual Behavior
Types of Individual Behavior

- **User-User** (link generation): befriending, sending a message, playing games, following, inviting
- **User-Community**: joining or leaving a community, participating in community discussions
- **User-Entity** (content generation): writing a post, posting a photo
I. Individual Behavior Analysis
Community Membership in Social Media

• Why do users join communities?
  – What factors affect the community-joining behavior of individuals?

• We can observe users who join communities and determine the factors that are common among them

• We require a population of users, a community C, and community membership information (i.e., users who are members of C).
  – Communities can be implicit: One can think of individuals buying a product as a community, and people buying the product for the first time as individuals joining the community.
  – To distinguish between users who have already joined the community and those who are now joining it, we need community memberships at two different times $t_1$ and $t_2$, with $t_2 > t_1$.
  – At $t_2$, we determine users such as $u$ who are currently members of the community, but were not members at $t_1$. These new users form the subpopulation that is analyzed for community-joining behavior.
Community Membership in Social Media

- Hypothesis: individuals are inclined toward an activity when their friends are engaged in the same activity.

- A factor that plays a role in users joining a community is the number of their friends who are already members of the community.

- In data mining terms, this translates to using the number of friends of an individual in a community as a feature to predict whether the individual joins the community (i.e., class attribute).
# Even More Features

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Feature</th>
</tr>
</thead>
</table>
| **Features related to the community, $C$.** (Edges between only members of the community are $E_C \subseteq E$.) | Number of members ($|C|$).
| | Number of individuals with a friend in $C$ (the fringe of $C$).
| | Number of edges with one end in the community and the other in the fringe.
| | Number of edges with both ends in the community, $|E_C|$.
| | The number of open triads: $|\{(u, v, w) | (u, v) \in E_C \land (v, w) \in E_C \land (u, w) \notin E_C \land u \neq w\}|$.
| | The number of closed triads: $|\{(u, v, w) | (u, v) \in E_C \land (v, w) \in E_C \land (u, w) \in E_C\}|$.
| | The ratio of closed to open triads.
| | The fraction of individuals in the fringe with at least $k$ friends in the community for $2 \leq k \leq 19$.
| | The number of posts and responses made by members of the community.
| | The number of members of the community with at least one post or response.
| | The number of responses per post. |

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Feature</th>
</tr>
</thead>
</table>
| **Features related to an individual $u$ and her set $S$ of friends in community $C$.** | Number of friends in community ($|S|$).
| | Number of adjacent pairs in $S$ ($|\{(u, v) | u, v \in S \land (u, v) \in E_C\}|$).
| | Number of pairs in $S$ connected via a path in $E_C$.
| | Average distance between friends connected via a path in $E_C$.
| | Number of community members reachable from $S$ using edges in $E_C$.
| | Average distance from $S$ to reachable community members using edges in $E_C$.
| | The number of posts and response made by individuals in $S$.
| | The number of individuals in $S$ with at least 1 post or response. |
Feature Importance Analysis

• which feature can help best determine whether individuals will join or not?

• We can use any feature selection algorithm. Feature selection algorithms OR

• We can use a classification algorithm, such as decision tree learning
  – Most important Features are Ranked Higher
Decimal Tree for Joining a Community

- Are these features well-designed?
  - We can evaluate using classification performance metrics.
Behavior Analysis Methodology

- **An observable behavior**: the behavior needs to be observable. E.g., accurately observing the joining of individuals (and possibly their joining times).

- **Features**: design relevant data features (covariates) that may or may not affect (or be affected by) the behavior – [we need a domain expert for this step]

- **Feature-Behavior Association**: Find the relationship between features and behavior. E.g., use decision Tree

- **Evaluation**: the findings are due to the features and not to externalities. E.g., use classification accuracy
  - We can also use randomization tests
  - Or causality testing algorithms
Granger Causality

Definition 10.1. Granger Causality. Assume we are given two temporal variables \( X = \{X_1, X_2, \ldots, X_t, X_{t+1}, \ldots\} \) and \( Y = \{Y_1, Y_2, \ldots, Y_t, Y_{t+1}, \ldots\} \). Variable \( X \) “Granger causes” variable \( Y \) when historical values of \( X \) can help better predict \( Y \) than just using the historical values of \( Y \).

• Consider a linear regression model
  – We can predict \( Y_{t+1} \) by using either \( Y_1, Y_2, \ldots, Y_t \) or a combination of \( X_1, X_2, \ldots, X_t \) and \( Y_1, Y_2, \ldots, Y_t \).

\[
Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \epsilon_1,
\]

\[
Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \sum_{i=1}^{t} b_i X_i + \epsilon_2,
\]

– If \( \epsilon_2 < \epsilon_1 \) then \( X \) Granger Causes \( Y \)
  • (why is this not causality)
II. Individual Behavior Modeling
Individual Behavior Modeling

• Models in Economics, Game Theory, and Network Science

• We can use:
  – **Threshold Models**: We need to learn thresholds and weights
  – Then, \( W_{ij} \) can be defined as the
    • fraction of times user \( i \) buys a product and user \( j \) buys the same product **soon** after that
    • When **soon**?
  – Similarly, thresholds can be estimated by taking into account the average number of friends who need to buy a product before user \( i \) decides to buy it.
  – What if no friends / or friends don’t buy the same products? Then we can find the most similar individuals or items (similar to collaborative filtering methods)
  – **Cascade Models**
III. Individual Behavior Prediction
Individual Behavior Prediction

• Most behaviors result in newly formed links in social media.
  – It can be a link to a user, as in befriending behavior;
  – A link to an entity, as in buying behavior; or
  – A link to a community, as in joining behavior.

• We can formulate many of these behaviors as a link prediction problem.
Link Prediction

• Most behaviors result in newly formed links in social media.
  – It can be a link to a user, as in befriending behavior;
  – A link to an entity, as in buying behavior; or
  – A link to a community, as in joining behavior.

• We can formulate many of these behaviors as a link prediction problem.
Link Prediction - Setup

- Link prediction assumes a graph $G(V; E)$.
- Let $e(u,v)$ represent an interaction (edge) between nodes $u$ and $v$, and let $t(e)$ denote the time of the interaction.
- Let $G[t_1,t_2]$ represent the subgraph of $G$ such that all edges are created between $t_1$ and $t_2$ — i.e., for all edges $e$ in this subgraph, $t_1 < t(e) < t_2$.
- Now given four time stamps $t_{11} < t_{12} < t_{21} < t_{22}$ a link prediction algorithm is given the subgraph $G[t_{11},t_{12}]$ (training interval) and is expected to predict edges in $G[t_{21},t_{22}]$ (testing interval).
- We can only predict edges for nodes that are present during the training period.
- Let $G(V_{train}; E_{train})$ be our training graph. Then, a link prediction algorithm generates a sorted list of most probable edges in $V_{train} \times V_{train} - E_{train}$.
• Assign $\sigma(x,y)$ to every edge $e(x,y)$
• Edges sorted by this value in decreasing order will create our ranked list of predictions
• Note that any similarity measure between two nodes can be used for link prediction; therefore, methods discussed in Chapter 3 are of practical use here.

• We will review some well-known methods
  – Node Neighborhood-Based Methods
  – Path-Based Methods
Computing Link Scores

- **Common Neighbors**: the more common neighbors that two nodes share, the more similar they are

\[ \sigma(x, y) = |N(x) \cap N(y)|. \]

- **Jaccard Similarity**: the likelihood of a node that is a neighbor of either \( x \) or \( y \) to be a common neighbor.

\[ \sigma(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}. \]
More Measures

- **Adamic-Adar**: if two individuals share a neighbor and that neighbor is a rare neighbor, it should have a higher impact on their similarity.

\[ \sigma(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log |N(z)|}. \]

- **Preferential Attachment**: nodes of higher degree have a higher chance of getting connected to incoming nodes

\[ \sigma(x, y) = |N(x)| \cdot |N(y)|. \]
Example

- Compute the edge score for edge (5,7)

\[
\begin{align*}
\text{(Common Neighbor) } \sigma(5,7) &= |\{4, 6\} \cap \{4\}| = 1 \\
\text{(Jaccard) } \sigma(5,7) &= \frac{|\{4, 6\} \cap \{4\}|}{|\{4, 6\} \cup \{4\}|} = \frac{1}{2} \\
\text{(Adamic and Adar) } \sigma(5,7) &= \frac{1}{\log |\{5, 6, 7\}|} = \frac{1}{\log 3} \\
\text{(Preferential Attachment) } \sigma(5,7) &= \frac{|\{4\}| \cdot |\{4, 6\}| = 1 \times 2 = 2}
\end{align*}
\]
Path-Based Measures

- Katz Measure:

\[ \sigma(x, y) = \sum_{l=1}^{\infty} \beta^l |\text{paths}_{x,y}^{<l>}|, \]

- \( |\text{paths}_{x,y}^{<l>}| \) denotes the number of paths of length \( l \) between \( x \) and \( y \). \( \beta \) is a constant that exponentially damps longer paths.
  - When small the measure reduces to common neighbor measure
  - It can be reformulated in Closed Form as

\[ (I - \beta A)^{-1} - I \]
Path-Based Measures

- **Hitting Time and Commute Time**: Consider a random walk that starts at node $x$ and moves to adjacent nodes uniformly. Hitting time is the expected number of random walk steps needed to reach $y$ starting from $x$.
  - A smaller hitting time implies a higher similarity; therefore, a negation can turn it into a similarity measure
  $$\sigma(x, y) = -H_{x,y}.$$  
  - If $y$ is highly connected random walks are more likely to visit $y$
    - We can normalize it using the stationary probability
      $$\sigma(x, y) = -H_{x,y} \pi_y.$$
• Hitting time is not symmetric, we can use commute time instead or its normalized version

\[ \sigma(x, y) = -(H_{x,y} + H_{y,x}). \]

\[ \sigma(x, y) = -(H_{x,y} \pi_y + H_{y,x} \pi_x). \]

• Some other examples
  – Rooted PageRank: the stationary probability of \( y \), when at each random walk run you can jump to \( x \) with probability \( P \) and to a random node with probability \( 1-P \)
  – SimRank: Recursive definition of similarity

\[ \sigma(x, y) = \gamma \cdot \frac{\sum_{x' \in \mathcal{N}(x)} \sum_{y' \in \mathcal{N}(y)} \sigma(x', y')}{|\mathcal{N}(x)||\mathcal{N}(y)|}, \]
• After one of the aforementioned measures is selected, a list of the top most similar pairs of nodes are selected.

• These pairs of nodes denote edges predicted to be the most likely to soon appear in the network.

• Performance (precision, recall, or accuracy) can be evaluated using the testing graph and by comparing the number of the testing graph’s edges that the link prediction algorithm successfully reveals.

• Performance is usually very low, since many edges are created due to reasons not solely available in a social network graph.

• So, a common baseline is to compare the performance with random edge predictors and report the factor improvements over random prediction.
Collective Behavior
Collective Behavior

• First Defined by sociologist Robert Park

• **Collective Behavior**: A group of individuals behaving in a similar way

• It can be planned and coordinated, but often is spontaneous and unplanned

• Examples:
  • Individuals standing in line for a new product release
  • Posting messages online to support a cause or to show support for an individual
I. Collective Behavior Analysis
Collective Behavior Analysis

• We can analyze collective behavior by analyzing individuals performing the behavior
• We can then put together the results of these analyses
  • The result would be the expected behavior for a large population

• OR, we can analyze the population as a whole
  • Not very popular for analysis, as individuals are ignored
  • Popular for Prediction purposes
Collective Behavior Analysis

- We can analyze collective behavior by analyzing individuals performing the behavior.
- We can then put together the results of these analyses.
  - The result would be the expected behavior for a large population.

- OR, we can analyze the population as a whole.
  - Not very popular for analysis, as individuals are ignored.
  - Popular for Prediction purposes.
Collective Behavior Analysis - Example

- Users migrate in social media due to their limited time and resources
- Sites are often interested in keeping their users, because they are valuable assets that help contribute to their growth and generate revenue by increasing traffic

- Two types of migrations:
  - **Site migration**: For any user who is a member of two sites \(s_1\) and \(s_2\) at time \(t_i\), and is only a member of \(s_2\) at time \(t_j > t_i\), then the user is said to have migrated from site \(s_1\) to site \(s_2\).
  - **Attention Migration**: For any user who is a member of two sites \(s_1\) and \(s_2\) and is active at both at time \(t_i\), if the user becomes inactive on \(s_1\) and remains active on \(s_2\) at time \(t_j > t_i\), then the user’s attention is said to have migrated away from site \(s_1\) and toward site \(s_2\).
Collective Behavior Analysis - Example

• Activity (or inactivity) of a user can be determined by observing the user’s actions performed on the site.

• We can consider a user active in interval \([t, t+X]\), if the user has performed at least one action on the site during this interval. Otherwise, the user is considered inactive.

• The interval could be measured at different granularity, such as days,
• weeks, months, and years.
  – It is common to set \(X = 1\) month.

• We can analyze migrations of individuals and then measure the rate at which the population of these individuals is migrating across sites.
  – We can use the methodology for individual behavior analysis
The Observable Behavior

• Sites migrating is rarely observed, while attention migration is clearly observable.

• We need to take multiple steps to observe it:
  • Users are required to be identified on multiple networks
  • Huan.liu1 on Facebook is Liuhuan on Twitter
Features

• **User Activity**: more active users are less likely to migrate
  • e.g., number of tweets, posts, photos

• **User Network Size**: a user with more social ties (i.e., friends) in a social network is less likely to move
  • e.g., number of friends

• **User Rank**: A user with high status in a network is less likely to move to a new one where he or she must spend more time getting established.
  • e.g., centrality scores
  • External rank: your citations, how many have referred to your article, ...
Feature-Behavior Association

• Given two snapshots of a network, we know if users migrated or not.

• Let vector \( Y \in \mathbb{R}^n \) indicate whether any of our \( n \) users have migrated or not.

• Let \( X_t \in \mathbb{R}^{3 \times n} \) be the features collected (activity, friends, rank) for any one of these users at time stamp \( t \).

• The correlation between features \( X \) and labels \( Y \) can be computed via logistic regression.

• How can we verify that this correlation is not random?
Evaluation Strategy

- To verify if the correlation between features and the migration behavior is not random
  - we can construct a random set of migrating users and compute $X_{\text{Random}}$ and $Y_{\text{Random}}$ for them as well
  - Find the correlation between these random variables (e.g., regression coefficients) and it should be significantly different from what we obtained using real-world observations
  - We can use Chi-square test for significance testing

$$
\chi^2 = \sum_{i=1}^{n} \frac{(A_i - R_i)^2}{R_i},
$$
II. Collective Behavior Modeling
Collective Behavior Modeling

• Collective behavior can be conveniently modeled using some of the techniques discussed in Chapter 4, “Network Models”.

• We want models that can mimic characteristics observable in the population.

• In network models, node properties rarely play a role; therefore, they are reasonable for modeling collective behavior.
III. Collective Behavior Prediction
Collective Behavior Prediction

- From previous chapters, we could use
  - Linear Influence Model (LIM)
  - Epidemic Models

- Collective behavior can be analyzed either in terms of individuals performing the collective behavior or based on the population as a whole.
  - It is more common to consider the population as a whole

- When predicting collective behavior, we are interested in predicting the intensity of a phenomenon, which is due to the collective behavior of the population (e.g., how many of them will vote?)

- We can utilize a data mining approach where features that describe the population well are used to predict a response variable (i.e., the intensity of the phenomenon).

- A training-testing framework or correlation analysis is used to determine the generalization and the accuracy of the predictions.
1. Set the target variable that is being predicted. In this case, it is the revenue that a movie produces.
   1. The revenue is the direct result of the collective behavior of going to the theater to watch the movie.
2. Determine the features in the population that may affect the target variable.
3. Predict the target variable using a supervised learning approach, utilizing the features determined in step 2.

   • the average hourly number of tweets related to the movie for each of the seven days
   • prior to the movie opening (seven features) and the number of opening theaters for the movie (one feature).
   • The predictions using this approach are closer to reality than that of the Hollywood Stock Exchange (HSX), which is the gold standard for predicting revenues for movies.
Generalizing the Idea

- **Target variable y**
  - Some feature $A$ that quantifies the attention
  - Some feature $P$ that quantifies the publicity

- Train a regression model

\[ y = w_1 A + w_2 P + \epsilon, \]