Social Network Analysis for Consumer Behavior Prediction

David Alfred Ostrowski
System Analytics
Research and Innovation Center
Ford Motor Company
dostrows@ford.com

Abstract

This paper examines the structural qualities of Social Networks towards the identification of trends. We consider a collection of Twitter messages towards the trending of technologically-related topics. For each individual topic, network cohesiveness is explored over time. We demonstrate that structural qualities reflecting community dynamics can provide insight to the prediction of long term trends. The goal of this work is to lend insight to the characterization of consumer behavior, particularly in the area of technology forecasting.

I. INTRODUCTION

Online Social Networks have increased substantially in recent years with the top Social Media sites acquiring hundreds of millions of users [1][2]. Through the increased accessibility of the internet, these outlets have encouraged users to interact and build relationships. Considered in depth by the sociological community, these relationships are closely compared to those found in real-life society [3]. As such, it has become increasingly important for organizations to understand these networks for the purpose of predictive analytics for internal as well as external application [4][5].

Such networks can be represented as graphs where nodes represent individuals and edges represent their relationships [6]. Given this network representation, prior work has demonstrated that the combined knowledge of community structure and relationship strength has important application to the area of web analytics [7]. In particular, when studied over time, this structural information can be informative towards suggesting what influence it will have on the future.

In the context of community as viewed from the perspective of a graph, there are qualities that are present among people who are highly influential which can be formally considered as power. Being a fundamental property of social structures, it has been suggested that power is relational [10][11]. Through research, a relationship has been established between power and density of a described structure [12].

Influence and power are strongly associated with transformation in the context of a community. Such a phenomenon is termed as emergence and is best described as a local action influencing an overall behavioral pattern. Through the examination of connections, research has also demonstrated that emergence can be objectified [12].

The study of such structural dynamics present strong support to what we know about the predictive nature of the internet [13][14][15]. Currently, a number of examples exist in the relation of trends as compared to the development of patterns [16][17][18]. Such patterns that exist around structures are highly indicative of the concept of communities [19]. These communities have been demonstrated across many studies as integral to the study of sustainability as well as economic opportunities.

The methodology outlined in this paper describes a means of examining the structural qualities within an online community. We apply this methodology to a collection of unstructured communication acquired from a web-based resource (Twitter) [20]. Our three primary inputs include topics of interest, web-based communications as well as the consideration of between whom the communication is directed. Building a graph-based structure based on communication, structural analysis is applied. From here, we examine the relationship over time of the
network structure to that of the interest of the concept.

The remainder of this paper is as follows: In Section Two we detail current work in the field. Section Three provides an examination of our methodology. Section Four presents our case studies and Section Five provides our conclusions.

II. CURRENT RESEARCH

Due to its constrained size and minimalistic attributes, a number of research efforts in Social Analytics have been directed at Twitter as a means of predicting trends. Among such efforts, (Achrekar et. al.) utilized Twitter to assist in the prediction of flu virus trends. This was accomplished by devising auto-regression models to predict data published by the Center of Disease Control (CDC) to the percentage of ‘visits’ to physicians [21]. (Iyengar et. al.) determined when an event started and ended by analyzing the content of Twitter messages using an SVM classifier and hidden Markov Model. Here, event boundaries in Twitter data were predicted for a set of events in the domains of sports, weather and social activities [22]. (Peng et. al.) demonstrated that conditional random fields could be used to improve prediction effectiveness by incorporating social relationships[23].

Focusing directly on structural qualities, (Gloor, Nann, Schoder) leveraged betweenness centrality of actors by weighing the context of their positions in a network. Through this application, they were able to extract and predict long term trends on the popularity of relevant concepts such as movies and politicians [7]. (Mislove et. al.) analyzed the structure of multiple online social networks. This involved the comparison of in-degree to out-degree representation though the investigation of side-free, power-law and small-world properties [24]. (Cantarane et. al.) applied social network techniques to analyze specific properties of social networking graphs. Among the qualities examined included centrality measurements, scaling laws and distribution of friendship [25].

A substantial amount of research from Link Mining has been leveraged to support Social Network concepts. These algorithms have supported performance among a number of activities including question answering, information retrieval and web-based data warehousing [26]. (Erbs et. al) demonstrated that data volume as well as training data is vital to link discovery with text-based approaches yielding superior results [27]. (Qian et. al) relied on the application of link mining to the Enron mail corpus to determine communities within linked nodes. Further work supported the identification of ‘common friends’ by relying on clustering. [28] Additional methods have explored link-predicates with applications among exploring data, distributed environments and spam analysis [29][30][31].

Research in Social Networks has also expanded on current search technology including PageRank and HITS [32][33][34]. Building on these methodologies, (Bharat, Henzinger and Chakraborti) proposed variations that exploit web page context to weight pages and links based on relevance.[35][36]. Relying on the topological structure of a graph (Sugiyama et. al.) successfully integrated several methods including network, quantitative, semantic, data processing, conversion and visualization-based components [37].

Investigations in Semantic Web technologies have also greatly affected development in Social Networks. Among them, (Zhou, Chen and Yu) integrated an ontology-based Social Network along with a statistical learning method towards Semantic Web data. This involved utilization of an extended FOAF (friend-of-a-friend) ontology applied as a mediation schema to integrate Social Networks and a hybrid entity reconciliation method to resolve entities of different data sources [38]. (Thushar and Thilagam) also relied on Semantic Web technology for the identification of associations between multiple domains within a Social Network [39].

A number of efforts in Relational Learning have supported Social Network analysis predicated on the concept of homophily-based associations to support learning. Among them include the application of probalistic modeling [40] collaborative relationship [41] and inference-based approaches [42].

Visualization techniques also provide for a substantial level of support in studying Social Networks. (Batajelj and Mrvar) developed tools for the visualization of large-scale networks allowing for support to identify vertex extracts and relations between clusters [43]. (Noel et. al.) calculated inter-item distances among combinations of elements from which hierarchical clustering dendograms are visualized to enhance measurement consistency between clusters and frequent itemsets [44]. (Levng et. al.) developed Social Viz which provided users with a means to view frequency relationships among multiple entities in a network [45].

III. METHODOLOGY

Our methodology consists of four key areas: network representation, tie strength, key players and cohesion [46]. To support the generation of a network, we first consider the application of basic data filtering techniques including matching on keywords and hashtags. After the data set is paired
down to a topic or set of topics we construct the networks. Here, all interactions may be considered. Frequently, this is addressed in the form of users who happen to 'follow' or 'like’ other users or their communications. A secondary means of edge generation is where a direct (personal) message exists between two entities. This exists in our test case in the form of a 'reply' to an existing post.

From our established network, tie strength is considered to evaluate strength/weakness of edges in our network, both as a means for graph reduction as well as clarity. Several methods are suitable among our data to serve as a proxy for the strength of a tie:

- Frequency of interaction
- Reciprocity in interaction
- Type of interaction (e.g. intimate or not)
- Relationship between entities (e.g friends, co-workers, relatives )
- Node Structure (neighborhood/relations/ affilliations)

Among the listed attributes, our main focus is to consider is frequency and reciprocity by generating a graph among entities (identifiers) who have replied to a message.

After consideration to our graph construction, we leverage the metric of centrality in understanding how key players interact in the graph. as well as their influence on the entire network. Centrality can be defined individually as a ratio of paths containing a point between two linking (separate) points. This can be formally expressed with \( b_{jk}(x_i) \) as the number of geodesics linking points \( x_j \) and \( x_k \) in a graph along with \( g_{jk}(x_i) \) where,

\[
b_{jk}(x_i) = \frac{g_{jk}(x_i)}{g_{jk}}
\]

is the proportion of geodesics linking \( x_j \) and \( x_k \) that contain \( x_i \).

To support the overall evaluation of a graph, an index of the centralization of a complete graph \( C_B \) is considered. It is based on the intuition that a graph is centralized to the degree that its communication flow is overwhelmingly dominated by a single point. The measure is defined as the average difference between the normalized centrality of the most central point \( C_B(p^*) \) and the normalized centralities of all other points in the graph.

Where,

\[
C_B'(x_i) = \frac{\sum_{i=1}^{n} [ C_B'(p^*) - C'(p_i) ]}{\text{n-1}}
\]

Finally, the networks structure is examined from the perspective of cohesion by the evaluation of connectiveness in a graph. Connectivenss is defined as the maximum number of elements which need to be removed to disconnect the remaining nodes from each other and is evaluated along Krackhardt's connectedness scoring. Existing as a means of expressing reachability in the termed graph, it represents the proportion of nodes (actors) that cannot be reached by other nodes. This is defined where a digraph G is equal to the fraction of all dyads \{i, j\} such that there exists an undirected path from i to j in G [48].

As opposed to comparing static metrics, both centrality and connectivity are evaluated over monthly intervals. Here the structural variation is considered as an indicator for changes in trends.

**IV. Case Study**

Our study consists of the application of stated methodologies towards a collection of 100M Twitter messages with time stamps ranging between 1/1/2010 and 6/31/2010 [49]. Our subject matter of interest was mobile Operating Systems. We focused on the three major competitive dominant (high presence) technologies – the Iphone, Blackberry and Android OS-based mobile devices. Figure one presents the market share for the three Operating Systems for the period between Jan 2010 and June 2010 [50].

![Figure 1. United States Market share for Mobile OS (Jan 2010 – Jun 2010)](image)

Each of these topics were identified by employing filtering on terms and phrases. The collection of messages were filtered on three topics and then separated by month. The volumes of filtered messages per each month are presented in Figure 2.

Within each data collection, we considered messages which were associated with a 'reply'
communication. As such the resulting adjacency metrics maintained at least one edge (minimally) between each node. To demonstrate the graphs, a reduced sample of 100 nodes from a selected month is presented for both the Android and Iphone collection. (Figures 3 and 4) Our approach in graph generation is reflected in this graphical presentation as each node has at least one connection with the Android sample presenting a higher degree of connectiveness and centrality.

Centrality was applied to view the properties of key players by employment of the Social Network Analysis ('R' programming) library [48]. The metric was considered for each topic across monthly intervals (Figure 5). This was examined over the entire sample for each month only including entities that had engaged in a reply conversation. Given our sample size, the metrics were very low as they ranged between .018 and .0122 among our three networks. This was in part due to our selected means of graph generation and secondly due to our sample size of approximately 1% of the entire Twitter feed (given the estimation of approximately 50M Twitter messages per day for the first and second quarters of 2010 as compared to our data set of 100M). Even though IOS maintained the largest overall sample size, the centrality measurement was the lowest, the graphs generated around the topic of Blackberry were highest corresponding to its gains in market share over the six month period.

Next, cohesion was examined, relying on the Krackhardt scoring mechanism for each monthly period as compared to each data set. Here, the Blackberry demonstrated a substantial increase corresponding to the US market share as determined earlier. Although maintaining the largest market share and sample size, the IOS maintained the lowest connectivity between all three ranging between 0.008 and 0.001.

Next, correlations were generated among the six sets of data, presented in figure 7, 8 and 9. First, the message sample sizes were compared against the market shares providing very little to negative correlation ranging between -.39 to .16. Corresponding to the relative volume collected, the IOS scored highest among centrality with the scores ranging between .71, .62, .85 for IOS, Android, and Blackberry. Although considering relatively moderate sample size, the BlackBerry metric scored extremely well in correlation with the market share. Connectedness was also directly correlated to the market share percentages with .91, .70 and .66 for
IOS, Android and Blackberry respectively. In spite of the relatively low connectivity metrics, IOS performed much higher than the other two mobile Operating Systems indicating that investigations in structural metrics as they change over time presents a substantial opportunity to understand their predictability.

Figure 6. Connectivity Measurements (Jan 2010 – Jun 2010)

<table>
<thead>
<tr>
<th>Volumes Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOS</td>
</tr>
<tr>
<td>Android</td>
</tr>
<tr>
<td>Blackberry</td>
</tr>
</tbody>
</table>

Figure 7. Volumes Correlation among our sample size. (Jan 2010 – Jun 2010)

<table>
<thead>
<tr>
<th>Centrality Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOS</td>
</tr>
<tr>
<td>Android</td>
</tr>
<tr>
<td>Blackberry</td>
</tr>
</tbody>
</table>

Figure 8. Correlation of Centrality metrics to Mobile OS marketshare (Jan 2010 – Jun 2010)

<table>
<thead>
<tr>
<th>Connectedness Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOS</td>
</tr>
<tr>
<td>Android</td>
</tr>
<tr>
<td>Blackberry</td>
</tr>
</tbody>
</table>

Figure 9. Correlation of Connectedness metrics to Mobile OS marketshare (Jan 2010 – Jun 2010)

V. CONCLUSION

We have considered the application of structural metrics to the trending of topics. This was accomplished by extracting topic-related data from a collection of 100M Twitter messages. Graphs were generated among the filtered collection by selecting only the messages that existed in the form of a reply. We then applied metrics to determine the effect of key players (Centrality) as well as metrics for cohesion (Connectedness). We were able to successfully correlate our graph structure metrics to the market share of three major mobile Operating Systems for the defined period. The disparity in correlation between volumes and our assigned structural metrics support that structure is a substantial indicator that is independent of changes in volumes within a given sample. As such, our metrics suggest that our Social Networks (utilized in real-time) could serve a means of predicting future consumer behavior.

Future work includes the incorporation of semantics on the filtering portion (graph generation) as well as reassigning weights to associated edges. Opportunities in this area also include examination of metrics (e.g. betweeness) among graphs integrating all three topics of comparison.

ACKNOWLEDGEMENT

Joseph Kelly, CEO of InfoChimps Corporation for Data Consultation and Support.

REFERENCES


[40] Achim Rettinger, Matthias Nickles, Volker Tresp: Statistical Relational Learning with Formal Ontologies, ECML PKDD '09 Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II


[42] Chunying Zhou; Huajun Chen; Tong Yu: Learning a Probabilistic Semantic Model from Heterogeneous Social Networks for Relationship Identification Tools with Artificial Intelligence, 2008. ICTAI ’08. 20th IEEE International Conference on Volume: 1


