Connectors, Mavens, Salesmen and More: An Actor-Based Online Social Network (OSN) Analysis Method Using Tensed Predicate Logic

Joshua S. White, PhD[1], Jeanna N. Matthews, PhD[2]
Clarkson University, Potsdam NY, 13676
[1]whitejs@clarkson.edu, [2]jnm@clarkson.edu

Abstract

In his bestselling book *"The Tipping Point"*, Malcolm Gladwell explored the "rules of epidemics" including how previously unpopular or unknown ideas can spread like viruses to reach a critical mass of people, and popularized three fundamental actor types that he argued are essential to the spread of ideas: connectors, mavens, and salesmen. In this paper, we supply some additional rigor to these popular definitions, allowing us to identify these actor types in online social networking data and study their impact. Specifically, we formally describe actor types using tensed predicate logic and apply these models to over 30 TB of data captured from the popular social networking service, Twitter. We also model additional actor types such as liaisons, bridges and stars as described by business organizational sociologists such as Allen T. Harrell, Arun Phadke, and James Thorp. We present our logical models both as logical predicates and as blocks of code written in RDF (Resource Description Framework) queries. These RDF queries can be applied directly to huge datasets such as the Twitter data stream and we present the initial results of doing so.

1 Introduction

Humans have been organizing themselves into social networks, both formally and informally, long before online social networks such as Twitter. In fact, we could conclude that social networks have existed as long as humans have shared thoughts and ideas with each other. However, it is increasingly clear that modern online social networks offer an unprecedented opportunity to study and quantify human interactions.

In this work, we were inspired by the qualitative descriptions of various key roles that individuals can play in the spread of information throughout social networks. For example, in Malcolm Gladwell's book, *"The Tipping Point"*, he describes a number of actor types such as mavens, connectors, and salesmen, that he argues are key to the spread of information [1]. Other actor types such as liaisons, bridges and stars have been described by business organizational sociologists such as Allen T. Harrell, Arun Phadke, and James Thorp [2, 3]. Our goal here is to translate these qualitative descriptions into formal logical models and then apply these models to a massive collection of actual data on the interaction of people and the spread of information through an online social networking service.

A multitude of methods for analyzing social networks have been developed over time. In the 1930's, social network analysis (SNA) was a type of research centered in the disciplines of sociology and psychology. By the 1950's, the fields of mathematics and statistics began contributing to formal models of SNA [4, 5]. More recently, the rise of large online social networks and the potential for analyzing the resulting big data sets, SNA has become a burgeoning research area for computer scientists as well.

We were also inspired by formal graph theoretical models in which nodes represent actors and the edges connecting these nodes represent the interactions between actors. Various types of analysis over these graphs are common including dyadic, triadic and group level analysis. Dyadic analysis looks at a single pair of actors and all of the links between them. Similarly, triadic analysis looks at a triple of actors and all of the links between them. Finally, group level analysis looks at items such as density, group centralization within the network, group and network diameter [6]. In this paper, we build on these mathematical formalisms and seek to automatically extract these formal models from online social networking data. In addition, we look to examine the dynamic nature of these graphs, how they change over time, and how information flows through them.

Another way to study the flow of information through online social networks is to follow the spread of specific memes. Memes are phenomena that manifest themselves in online cultural environments like Tumblr, 4Chan, Twitter, Facebook and other social media networks. They are often represented in a very precise and searchable way such as with the use of specific designators (hashtag-word,

for example "#kony2012"). By following where specific memes first appear, how they spread in popularity, and reach throughout a social network, we can map to specific ideas about how information spreads. For example, we can quantify the impact of actors such as those popularized by Gladwell in *The Tipping Point*.

2 Online Social Networks

In this work, we focus on Twitter in particular, but there are many kinds of online social networking services and our analysis methods could be applied to any of these types. A commonly accepted classification of various social media networks has been proposed in a number of places inlcuding a document called "Publicly Available Social Media Monitoring and Situational Awareness Initiative" published by the Department for Homeland Security in 2010 [7]. Important categories of online social networking services include journal sites, profile-sharing sites, recommendation sites, multimedia sites and mapping sites.

Journal sites such as blogging and micro-blogging sites are the most proliferate and hardest to track, these include everything from individual personal blogs to mass micro blog aggregation sites, such as Twitter. Profile-Sharing sites encourage users to share personal information with individuals they know, sometimes only casually. Facebook is the largest example of this type. Recomendation sites offer general search and browsing results based on recommendations or suggested content types by other users. Examples include shopping sites like Amazon. Multimedia sites allow users to share their own content, video, audio and images, for others to review. The most popular of these sites are Youtube, Flickr, Vimeo and Hulu. Mapping sites offering geographic mapping services allow users to share maps about things going on in their communities on a wide range of topics, from politics to health. One such example is Google Flu, which tracks user reports of flu outbreaks.

All these types of online social media networks allow users to contribute to the site's content and most include the ability for a user to associate themselves with other users in a specific formalized relationship. For different online social networks, the nature of the relationships can be different (e.g. friending in Facebook is not the same relationship as following in Twitter) and thus the meaning of each link or relationship can be subtly different. While Twitter does support following and being followed by, this is not the same as the so called "friending" that occurs on other networks because it is not reciprocal (i.e. Facebook "friending" requires that both parties accept the associattion while Twitter's following relationships only requires the action of one party). In Twitter, another important indication of relationship is when a user who has a lot messages directed to them directly (@username). We use actions of this type to indicate relationships between users or actors so that we can study the characteristics of the underlying social network.

What constitutes as a link varies dependent on the type of online social network being studied. For instance we consider that a link that exists between the end-user and the person who posted an article on a recommendation site as *established* if the article was clicked on. In a weighted system this might be considered a tier one link establishment, or the lowest edge weight which signifies little connectivity. In these same networks clicking on an article posters profile might constitute a tier two link establishment. Since the common component of the network is the article poster, anyone who clicks on that persons profile or article they posted are all part of a single sub-network. While in this work we have focused on microblogging sites like Twitter, we note that our actor description logic can be used on any link establishment type.

Online social networks offer many unique advantages for studying the flow of information in human social networks. Specifically, online social networks:

- **Generate massive amounts of hard data on social interactions.** Online social networks capture each interaction in a precise, digital representation that can be stored, queried and processed. This is true of all online social networks, but for researchers, sites like Twitter that offer publicly available access to the data posted by users are especially attractive. In Twitter, any user can "follow" any other user. Online social networking sites like Facebook where the data is not publicly available could be studied in a similar way, but only by researchers with privileged access to the collected data and with the informed consent of the users whose data is being studied.
- **Are becoming increasingly ubiquitous.** The most widely known online social networks have become truly ubiquitous in many parts of the world. In 2013, there were 554,750,000 reported Twitter users, and 1,110,000,000 reported Facebook users [15, 17]. "Following", "Friending" and "Liking" have become normal parts of our language, much as "Googling" has become synonymous with searching. Online social networks have increased in popularity over the last few years, so much so, that 80% of North Americans and 60% of adults worldwide use them [21].
- **Offer a rich representation of a wide range of human interactions.** Online social networks are not just for updating social statuses. They increasingly represent a substantive reflection of society as a whole. For example, we see evidence of online social networks as agents of change [8, 9]. Social networking was credited with helping to influence the thoughts and actions of large groups in movements such as the Arab Spring [12, 13].

In another example, the darker side of society is clearly represented in online social networking including fraud, deception, phishing and other forms of abuse [10, 11].

- **Facilitate increased interaction.** In many ways, these networks enable increased communication with friends and family all over the world through shared images, videos, and messages in a way that was not possible just a decade ago. The identities cultivated on these sites have become so important to some, that they can not imagine a life without them [22]. With access available 24/7, these networks are becoming a standard for establishing new relationships and solidifying old ones.

3 Formalizing Actor Types with Tensed Predicate Logic

In this section, we present formal models of each actor type. Specifically, we characterize each actor type using tensed predicate logic and provide a graph representation of the actor types to illustrate the key relationships. In the next section, we will translate these models into semantic web queries or blocks of code written in RDF that can used to directly process terabytes of data collected from online social networks.

Actors in a social network context are simply individuals or groups of individuals. In the online social network case, actors are typically individual accounts in an online social networking service and are represented as nodes in a social networking graph. We build a formal graph of nodes and links between these nodes based on various types of interactions. Groups of individuals can be identified and their interactions studied. For example, dyads are simply two individual actors and triads are three actors. Larger groups are sometimes identified as organizations. One of the foundantional classification of actors was proposed by Allen Harrell [2]. He proposed three types including stars, bridges, and liaisons. We extend this by considering those presented in Gladwell's popular book *"The Tipping Point"*: which include connectors, mavens, and salesman [1].

First, we translate the qualitiative description of each actor type into a formal model. Certainly, some traits are harder than others to measure or quantify in an online social network (e.g. charisma or confidence), but many traits have natural representations in online social network data. Table 1 introduces some notation and Table 2 presents the tensed predicate logic definitions of each of the actor types we are considering - isolate, star, bridge, liason, maven and salesman. In this section, we discuss these actor types in more detail including why we classify star, bridge and liason as subtypes of connector and why we distinguish between prospectively and retrospectively identified liasons.

A key aspect of our work on actor description logic is our attempt to factor in the order of events, known as temporal logic. The definition for temporal logic is broad and is a blanket term for all logic frameworks that take time into account [29]. For our purposes, as is shown in Table 1, tense predicate logic combines truth-functional operators of propositional and predicate calculus, quantification of predicate calculus, and modal operators of Prior's tense logic [19]. Time components are defined as $(\mathbf{P}, \mathbf{F}, \mathbf{H}, \mathbf{G}.)$ In addition, we have added terms from graph analysis such as node and edge along with representations for special objects, such as messages and centrality.

Table 1: Tensed Predicate Logic

Symbol	Interpretation
Tensed Predicate Logic	
$\neg$	Negation
$\wedge$	Conjunction
$\vee$	Disjunction
$\rightarrow$	Conditional
$\leftrightarrow$	Biconditional
$\forall$	Universal Quantification
$\exists$	Existential Quantification
$\mathbf{P}$	It has at some time been the case that ...
$\mathbf{F}$	It will at some time be the case that ...
$\mathbf{H}$	It has always been the case that ...
$\mathbf{G}$	It will always be the case that ...
Custom Terms	
$edge(x,y)$	there is an edge from x to y
$edge'(x,y)$	there is an edge from x to y or from y to x
$group(x)$	is a group of nodes (actors)
$node(x)$	is a single node (actor) within the network
$network(x)$	is a network graph consisting of all nodes (actors) within
msg	is the uniquely identifiable contents of a communication (message)
$cent(i)$	is the measure of Betweenness Centrality for a node (actor) in the network as calculated by: $C_B(i) = \sum_{j<k} \left(g_{jk}(i)/g_{jk}\right)$

The most relevant related work is *Where the Blogs Tip: Connectors, Mavens, Salesmen and Translators of the Blogosphere* [40] by Budak, et. al. In this work, the authors investigated whether or not Gladwell's actor types exist in the so-called blogosphere. They formally defined Gladwell's actor types and attempted a study of their effect on successful Internet campaigns. Additionally, Budak, et. al. applied their own modified interpretation of Gladwell's actor descriptions, by implementing a graph analysis approach. For instance, a connector is defined in graph analysis as a node with a high degree of centrality within a subgraph. Such representations are more complicated for other actor types, as they must try to correlate influence based on actual success or failure, not simply potential for

success.

To address these problems, the logical descriptions that we propose here take time into account for each actor type. Time is important when defining some actor types, because it allows us to track the order in which links are established rather than simply the set of links that exist at one point in time or ever exisit.

In addition to the definitions in Table 1, we also translate our tensed predicate logic definitions into semantic web queries. These queries take less time to process in many cases than prior approaches based solely on graph analysis approaches. We present example RDF implementations in Section 4.

3.1 Isolates

Isolates, a derivative of developmental psychology, are asocial actor types [27]. It describes members of a study-group who are unconnected to any group. We show an isolate (e) in Figure 1 as not connected or communicating with the rest of the group.

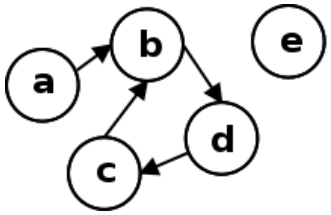

Figure 1: Graph Representation of an Isolate (e).

An isolate for the sake of our analysis is defined as a person who does not direct their messages at anyone in particular and does not repeat anyone's sentiment directly, ie: "retweet". In the case of Figure 1 we say that actor (e) might be an isolate for Graph (a,b,c,d,e), if at no point in the past or future of the network, does (e) communicate with any other actor within said network, as is shown in Equation (1). When needed, we can generalize this further by creating wildcards for unknown potential connecting nodes. However, for simplicity sake, we use only specific node lettering in the figure.

3.2 Connectors

The Connector archetype is defined as a person that is quick to search their own knowledge base, or their connections, for knowledge. Connectors have a large number of connections and are willing to share these connections, thus becoming social bridges. These individuals are confident, energetic, social, and have the innate ability to befriend people with a wide range of views.

A connector on social sites, would have a lot of "friends". On reciprocal social networks they would have a lot of messages directed at, (@username), them directly. We note that these traits can be hard to measure in a system such as Twitter and in this section, we discuss how we can identify connectors in our example Twitter data set. We further subdivide connectors into three groups for the sake of more comprehensive data analysis: bridges, liaisons, and stars. Each of these subclasses will be discussed in more detail.

3.2.1 Stars

Stars are actors within a particular group with the largest number of percentage based interactions. This term was first used in "An Experimental Study of the Small World Problem" [26] in 1969. The small-world experiment was actually a series of experiments conducted in the US by Travers and Milgram to examined the average length of a connection path in a social network. This series of experiments suggested that from a societal view, a person is connected to every other person on the planet through very short connection paths. This work was later referred to as "the principle of six degrees of separation."

We say that a star a node with the shortest paths between the majority of actors in a Group $(G_{(a)})$, shown as (b) in Figure 2. This is known as a measure of betweenness centrality when applied to the network as a whole. The node with the highest betweenness centrality value is considered the most central, or the shortest path between the majority of nodes. We apply this to the subgraph of a single group in Equation (2)

3.2.2 Bridges

We consider bridges to be another sub-class of the connector actor type. Their purpose is different than that of a liaison. A bridge is a type of actor that has relationships outside of a focal group. He/she then connects that focal group to another actor or group. Bridges, unlike stars, have weak network ties, usually only two connections, however they provide the shortest path between two distinct groups or individuals as represented by $(N_{(a)})$ in Figure 3.

We say that an actor is acting as a bridge when it is weakly connected, and yet has a high betweenness centrality in the total graph while at the same time connecting two or more groups. Equation (3) states that bridge (b) exists if no edges exist directly between the two different groups AND an edge between (b) and both nodes (c) and (e) AND the centrality of (b) is higher than the centrality of (c) and (e).

3.2.3 Liaisons

Liaisons are considered the primary sub-class of the connector type. They link many groups together through

Table 2: Actor Type Logics

Actor Type	Logic	
Isolate	$\forall a\,[Isolate(a) \leftrightarrow \mathbf{G}\,[\forall b \neg edge(a,b)]]$	(1)
Connector: Star	$\forall a\,\Big(Star(a) \leftrightarrow \neg\exists b\,\Big(cent(b) > cent(a)\Big)\Big)$	(2)
Connector: Bridge	$\forall b\,\left(Bridge(b) \leftrightarrow \exists c,e\,\left(\begin{array}{c} c \neq e \wedge edge'(b,c) \wedge edge'(b,e) \wedge \\ \forall x\,\big(edge'(b,x) \rightarrow (x = c \vee x = e)\big) \wedge \\ cent(b) > cent(c) \wedge cent(b) > cent(e) \end{array}\right)\right)$	(3)
Connector: Liaison (Prospective)	$\forall a,b,c\,(Liaison(a,b,c) \leftrightarrow \mathbf{F}\,(edge(a,b) \wedge \mathbf{F}\,(edge(b,c) \wedge \mathbf{F}\,(edge(c,a) \wedge \mathbf{H}\neg edge(c,a)))))$	(4)
Connector: Liaison (Retrospective)	$\forall a,b,c\,(Liaison(a,b,c) \leftrightarrow \mathbf{P}\,(edge(c,a) \wedge \mathbf{H}\neg edge(c,a) \wedge \mathbf{P}\,(edge(b,c) \wedge \mathbf{P}edge(a,b))))$	(5)
Maven	$\forall m\,(Maven(m) \leftrightarrow \exists i,g\,\mathbf{F}\,(edge(i,m,msg) \wedge \mathbf{F}\,(edge(g,m) \wedge \mathbf{F}\,(edge(m,g,msg)))))$	(6)
Salesman	$\forall s\,(Salesman(s) \leftrightarrow \exists i,g\,\mathbf{F}\,(edge(i,s,msg) \wedge \mathbf{F}\,(edge(s,g,msg) \wedge \mathbf{H}\neg edge(g,s))))$	(7)

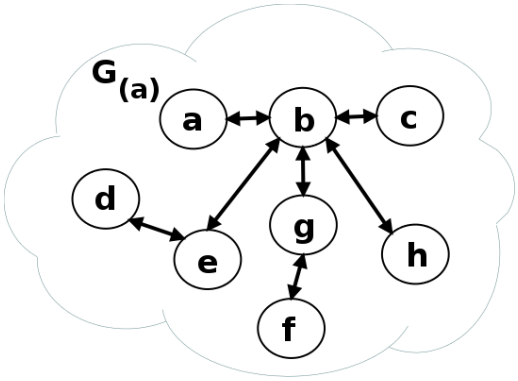

Figure 2: Graph Representation of a Star (b) within a Group ($G_{(a)}$)

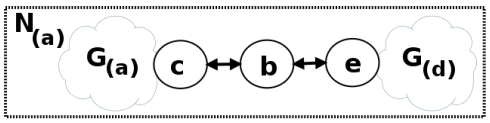

Figure 3: Graph Representation of Bridge (b) connecting two groups ($G_{(a)}$) and ($G_{(d)}$)

their individual connections. This actor type provides the shortest path between groups. Liaisons can also be described as individuals who introduce people or setup connections on the behalf of another. Notice that the order of events is important, (a) connects to (b), then (b) connects to (c), then (c) connects to a for the first time. In other words, (b) introduces (c) to (a). We illustrate this definition in Figure 4. We take into account that the relationship that takes place in Figure 4 does so over a wide period of time and each step may occur close to or significantly later than the previous.

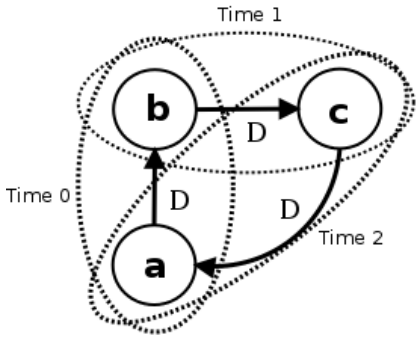

Figure 4: Graph Representation of a Liaison Between (a) and (c)

In Figure 4 (a) needs information from an unknown source, represented as (c). (a) communicates this need to (b), who then connects to (c) in an attempt to aid (a), thus connecting (a) to (c). This example can be used for creating an actor definition for both a static-historical dataset and a live one.

We make a distiction between liaisons that are identi-

fied prospectively versus retrospectively. If we consider prospectively that data about an actor has yet to occur, we can say that actor (*b*) will be a *Liaison* for *c* and *a* as shown in Equation (4). If we redefine *Liaison* as in this way, for use on a static-historical dataset, we would change the first **F** from Table 1 to *"it will some time be the case that"* and in Equation (4) to a **P** *"it has at some time been the case that"* , in which case (*b*) is a *Liaison* if previous communication with (*a*) had occurred. This representation allows the initial communication to have occurred in the past. This does not suggest that subsequent communications can not still occur in the future. A pure retrospective representation would consider the events in reverse chronological order. This is shown in Equation (5), where (*b*) is a *Liaison*, if (*c*) communicates with (*a*) for the first time after communication from (*b*), after (*b*) receives communication from (*a*). This distinction is especially important when considering the differences between analysis on a static dataset, versus a dataset that is growing in real time.

3.3 Mavens

Mavens are individuals who others rely on for new information. A maven is an actor who collects new knowledge, usually from other individuals and is willing to share it when asked. Mavens gather information, many times in an effort to solve a problem that they themselves are having. However, due to their need to share and their better than average social and communication skills, they are able to efficiently pass on the knowledge they have collected. A maven will generally not stop at passing this knowledge on to just one person. They usually will repeat it over and over again to many individuals, until it becomes a social epidemic. Gladwell notes that mavens can be thought of as information brokers, since they often trade or ask for more information in return [1].

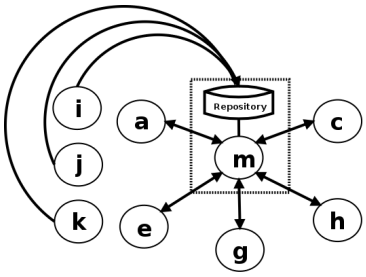

Figure 5: Graph Representation of a maven (m) receiving information from actors (i, j, k) storing it and later disseminating it upon request

In essence, a maven is an actor that provides information to their connections upon request from a repository they have assembled from others. These bidirectional communications are shown in Figure 5 as lines which connect actor (m) to other actors with arrows on both ends. Equation (6) introduces the message operator (*msg*) into Table 1 to handle the representation of passing unique pieces of information.

3.4 Salesmen

Salesmen are described as charismatic persuaders who have the ability to convince others to agree with them. Figure 6 describing a salesman is similar to the previous Figure 5 describing a maven, but with two key differences. First, Figure 6 shows that communications from actor (s) is outward uni-directional. This is because a salesman is marketing rather than collecting information. Information is also not being requested. Second, it shows the salesman pulling information from a repository, represented in Equation (7) as (i), with less emphasis on the repository being filled from multiple sources. From the salesman's perspective, the repository could consist of only a single source, such as a maven. The repository is typically filled with curated information as opposed to a maven, who serves as the curator.

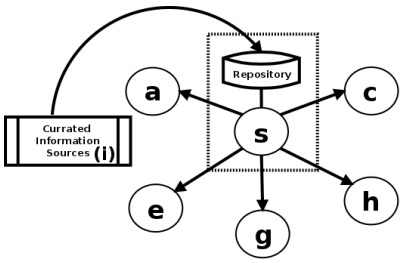

Figure 6: Graph representation of a salesman (s) receiving information from curated sources such as businesses or marketing literature and disseminating it in mass

4 Translating Formal Definitions into Semantic Web Queries

In the last section, we presented formal models of our actor types as tensed predicate logic and illustrated them with graphical models. In this section, we translate these models into semantic web queries or blocks of code written in RDF. Queries like this can be used to directly run queries on the data collected from online social networks.

Efficient queries are essential because the analysis of online social networks often requires processing of large data sets to which new data is constantly being added [23]. Users on sites like Twitter can produce more than 58 million tweets per day. Facebook users often produce more than 1 billion posts per day, with an average useful lifespan of only 3 hours each [25].

There are various ontology languages for the semantic web. OWL, the Web Ontology Language is perhaps the most widely used. We concentrate here on querying using RDFLib, a pure Python interface for dealing with RDF (Resource Description Framework). [32]. Both RDFLib and OWL provide near and long term solutions for our analysis by supporting both current logical analysis and any future semantic analysis that we may include. This is done by standardizing on SPARQL as a query language. RDFLib makes this easy by simply passing a standard query as a string.

We do not have space to include the full RDF query for each actor type. Instead, we chose one actor type and show an end to end example of translating its formal definition into a semantic web query and running that query on a large set of collected Twitter data.

4.1 Sample RDF Data and Query

Code Block 1 contains the RDF data for the liason actor type. It makes a good example because the liason actor type contains some of the most complex actor relationships.[1]

In Code Block 1, there are three users listed in the data: (a) user1, (b) user2, and (c) user3. As was true before in Figure 4, (b) is the actor that we refer to as a liaison.

Code Block 1: Example RDF Data for liaison querying

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:j.0="urn:" >
  <rdf:Description rdf:about="urn:communication100">
    <j.0:hour rdf:datatype="http://www.w3.org/2001/XMLSchema#long">2012022001</j.0:hour>
    <j.0:target rdf:resource="urn:user2"/>
    <j.0:source rdf:resource="urn:user1"/>
  </rdf:Description>
  <rdf:Description rdf:about="urn:communication101">
    <j.0:hour rdf:datatype="http://www.w3.org/2001/XMLSchema#long">2012022002</j.0:hour>
    <j.0:target rdf:resource="urn:user3"/>
    <j.0:source rdf:resource="urn:user2"/>
  </rdf:Description>
    <rdf:Description rdf:about="urn:communication102">
    <j.0:hour rdf:datatype="http://www.w3.org/2001/XMLSchema#long">2012022003</j.0:hour>
    <j.0:target rdf:resource="urn:user1"/>
    <j.0:source rdf:resource="urn:user3"/>
  </rdf:Description>
</rdf:RDF>
```

[1]However, one interesting thing not illustrated with the liason type is that some actor types require additional graph based input such as measures of centrality. SPARQL can not calculate things like in-degree, out-degree, or measures of centrality. However, systems such as SemSNA (Semantics for Social Network Analysis) extend basic query capability by providing an ontology to describe social network specific features [48].

We translate the description of a liaison and the logical representation put forth in Equation (4) and represent it as a SPARQL query in Code Block 2. This query states that we expect to see at least three communications, in this case graph edges, occur: (a,b), (b,c), and (c,a). Our first filter statement makes sure that each user has a unique name. The second filter states the order of relationship occurrence using the included timestamps. In this case (a,b) must come before (b,c) which must come before (c,a). Finally to remove any ambiguity we create one last slightly more complex filter statement to make sure that no previous (c,a) communication has taken place before the liaison event.

Using RDFLib, we run the query listed in Code Block 2 against the dataset shown in Code Block 1, the formatted output is shown in Code Block 3. We can see that three district user-names were found: (a) user1, (b) user2, and (c) user3. We can see that three edges were constructed in the graph following the (a,b) then (a,c) then (c,a) order based on their respective time-stamps. We state that no (c,a) relationship occurred before the initial (a,b) edge was constructed. Finally we conclude that (b) is the liaison actor type.

Code Block 2: Example RDF Query for a liaison in Python

```
Qliaison = g.query("""BASE <http://xmlns.com/foaf/0.1/>
    PREFIX : <urn:>
    SELECT ?a ?b ?c ?t0 ?t1 ?t2
    WHERE {
        [ :source ?a ; :target ?b ; :hour ?t0 ] .
        [ :source ?b ; :target ?c ; :hour ?t1 ] .
        [ :source ?c ; :target ?a ; :hour ?t2 ] .
        FILTER( ?a != ?b && ?b != ?c )
        FILTER( ?t0 < ?t1 && ?t1 < ?t2 )
        FILTER( NOT EXISTS {
        [ :source ?c ; :target ?a ; :hour ?tx ] .
        FILTER( ?tx < ?t0 || ?tx < ?t1 ) })}""")
```

Code Block 3: Example output of an RDF query (Shown in Code Block 2) for a liaison given the example dataset (Shown in Code Block 1)

```
   (a)     (b)     (c)
==========================
| user1 | user2 | user3  |
==========================

  Slot     Edge          Hour
====================================
|  t0  | (a) --> (b) | 2012022001  |
|  t1  | (b) --> (c) | 2012022002  |
|  t2  | (c) --> (a) | 2012022003  |
====================================

No previous (c) --> (a) relationship was found before time-stamp 2012022001

We conclude (b) user2 is a Liaison
```

4.2 Sample Results

Over the course of 2012, we collected multiple Terabytes of data from Twitter. Recently released estimates place total Twitter traffic at 175 million tweets per day [20]. Comparing the amount of data we collected daily to Twitter's own reports on traffic per day, we estimate that we collected between 50% and 80% of all Twitter traffic. Our complete 2012 dataset consists of 147 days of Twitter data, stored as 30 Terabytes of gzip compressed JSON formatted data. A typical day's compressed data can vary significantly, but averages at 70 GB.

For the results in this section, we chose to focus on 31 days of Twitter data from February 20th to March 20th of 2012 that allow us to study the flow of a particular topic or meme through a social network. In particular, February 20 2012 was the release of the KONY2012 video on Vimeo [33]. This video sparked a very interesting case of an information epidemic. Before the release of this video, many people were unaware of Joseph Kony or the situation in Uganda. Awareness spread quickly through online social networks and even crossed over in the mainstream media. Identifying the actor types involved in this information epidemic is a perfect application for our actor characterizations.

We first filtered the tweets captured during this time period looking for those that were directed at specific users (@username) and which contained a KONY2012 related hashtag such as: #KONY2012 or #StopKONY. We considered these messages to be intentionally focused between one user and another. We translated this into RDF data and processed against our RDF queries for the 4 unique actor types as shown in Table 3 .

Table 3: KONY2012 Related Directed Tweet Actor Query Results

Query	Number of Records
Edges	1,070,910
Isolates	48,060
Liaisons	37,530
Mavens	1,790
Salesmen	391

The output data from our initial MapReduce job was a 48MB CSV file. This file contained columns for: time, source, target, and message. After conversion to RDF, the file size ballooned to 450MB. Processing times for straight RDF versus serializing RDF into our SPARQL query methods varied. Table 4 shows the time results using the liaison query shown in Table 4. These tests were done on an "AMD FX(tm)-8120 Eight-Core Processor" with 24GB of DDR-3 memory.

In an effort to understand the load our system might be under and better plan for a future system, we tested the configurations shown in Table 4. First, we converted our existing node relationship CSV file to RDF using Python. This process took only 18 seconds. Given the number of edges contained within the CSV, 1,070,910 in total, this number of 18 seconds was a promising start. Processing the RDF file with Apache Jena, a system for running SPARQL queries efficiently against RDF [31], took a total of 6.285 minutes. This process was load intensive, it kept our eight-core processor running at maximum capacity for the duration of the run time.

Table 4: KONY2012 Related Directed Tweets - Liaison Query Result Transaction Speeds

Approach	Time
Conversion of CSV to RDF using Python	18 sec
RDF file procd. w/Jena (8 thr.)	6.285 min
RDF file procd. w/RDFLib (1 thr.)	13.151 hr
RDF file procd. w/RDFLib (8 thr.)	35.854 min
Serialized CSV–RDF procd. w/RDFLib (1 thr.)	13.159 hr
Serialized CSV–RDF procd. w/RDFLib (8 thr.)	36.762 min

RDFLib has the same capability to apply SPARQL queries to an RDF dataset. These queries are simply wrapped in a string, as shown in Code Block 2. This allowed us to easily control the number of threads that would be assigned to each particular query. Processing the same RDF file with RDFLib as a single thread, took an astounding 13.151 hours. While the increase in time was expected, this amount of increase was not. We tested RDFLib running at 8 threads and found that it took a more reasonable 35.854 minutes to process the dataset. This run took longer than Jena's run but was still reasonable.

5 Related Work

The social sciences have studied traditional social networks for some time. There are a myriad of books and papers on the topic. Some classics are Milgram's *"The Small World Problem"* [26], Wasserman's *"Social Network Analysis: Methods and Applications"* [6], and Pool's, et. al *"Contracts and Influence: Social Networks"* [3].

Wasserman's work, published in 1994, provides a textbook-like approach to social network analysis from a mathematical and combinatorics approach. Wasserman presents each topic clearly, but without consideration of computational feasibility as it relates to large scale SNA. It is a foundational work that presents an excellent overview of the topic.

Milgram's work, presented in 1967, is a groundbreaking paper on what he termed the small world problem. Milgram provides evidence that the average path length

between any two people in the United States consists of six hops. This is often referred to as *"six degrees of separation"* [35].

Milgrams work was extended by Pool, et. al in their 1978 paper on contracts and influence within social networks. These authors argued that social networks can be broken down into smaller parts. They categorize individuals according to whether they have strong or weak ties to the network. This work was the basis of modern clustering approaches that have been built into most SNA tools.

Since the advent of online social networks, much of the social science and mathematics work previously developed for traditional SNA has been tested and retested. Adamic's et al. research entitled *"A Social Network Caught in the Web"* [36], presented a study of early online social networks at Stanford. They found that the networks studied had many small world characteristics and clustering coefficients that were predicted in Milgram and Pool's research.

Further work by Liben, et al., published in 2005, entitled *"Geographic Routing in Social Networks."* [37], showed that there existed a strong correlation of friendship and geographic location of users on the LiveJournal network. This study was essentially reproduced by Kumar, et al. in their work *"Structure and Evolution of Online Social Networks"* in 2006 [38]. Kumar, et al. used two different social networks, both hosted by Yahoo and found that they both had a significant social clustering coefficient. Again, this was validated by Girvan, et al. in *"Community Structure in Social and Biological Networks"* [39] that helped to validate that online social network users form tight groups which can be seen through the calculation of these social clustering coefficients.

As previously discussed, Gladwell's work focused on the social characteristics of individuals that influence a tipping point in some way. Gladwell's work has been validated by some and shunned by others. For example one opponent to the ideas put forth in Gladwell's work, Levitt, argues in his book *"Freakonomics"* [41] that some of the statements Gladwell made about who is and isn't influential are simply wrong. The case that Levitt is most focused on, is Gladwell's discussion of the decrease in NYC crime rates as a result of the NYPD's *fixing of broken windows*. This theory was based on the idea that the establishment of environmental norms, like keeping streets clean, removing graffiti, and fixing broken windows, leads to a reduction in crime [42].

Levitt states that multiple factors cause change as opposed to Gladwell who believes in a tipping point and the role of key individuals in producing the factors that lead to these tipping points. Levitt states that the decrease in NYC crime was a result of multiple factors including the reduction of abandoned children due to Roe vs. Wade as opposed to the actions of just the police department [43]. His argument is that when abortion was made legal, there were far fewer homeless children, which reduced crime all over the country, not just in NYC. Gladwell's work is further attacked by Jonah Berger in his work *"Contagious: Why Things Catch On"* [44]. Berger states that influentials like mavens simply don't exist. He states that it is not the few that cause an idea to spread, but instead how well the idea is presented.

We agree that there are numerous factors that cause change to occur within a society however, but we also support Gladwell's idea that individuals do indeed lead to the creation of tipping points. Due to the redundancy of other works that have focused on the centrality of individuals within a network, their physical location and on key individual's personality traits, we preferred to formalize the logic behind actor descriptions in the hopes that this will be used in further research to either support or disprove the role of key individuals in the tipping point theory. In the future, we plan to apply our actor descriptions to the rise and fall of popular memes in Twitter and quantify the impact of actor types identified. We believe this will offer a unique opportunity to support or refute ideas such as those presented in *"The Tipping Point"*.

In Budak, et al. 2010 paper, entitled *"Where the Blogs Tip: Connectors, Mavens, Salesmen and Translators of the Blogosphere"* [40] the authors attempt to justify Gladwell's work. They state that the fundamental function of connectors, mavens, and salesman and instrumental actor types in these online social networks is fundamental to SNA. In their paper a formal definition for each of these actor types is presented. The authors go on to present a new class of actor that they term a Translator, who serves as a bridge between different interest groups.

Budak, et al. make some potentially incorrect assumptions regarding personality traits, in order to better "fit" their models. For instance, the authors consider the actor type connector to be defined only by its centrality within the graph. They do not take into account the directionality of these connections, as well as the order of occurrence over time. Their definitions were therefore not used within this work. While clustering coefficients are useful, they too were not used in this body of work due to lack of scalability. We believe the equations we have created can be more simplistically represented in a distributed processing environment, such as MapReduce.

Renfro in his 2001 thesis entitled *"Modeling and Analysis of Social Networks"* [45], presented a method for developing measurements of interpersonal influence. He used these measurements and tested various other models in the study of the *Organizational Phenomena*. His results were incorporated in Clark's thesis in 2005 entitled *"Modeling and Analysis of Clandestine Networks"* [46] which presented improvements to Renfro's metrics that took into account the existence of uncooperative network models. Hamill evolved these metrics in his 2006 thesis, *"Analy-*

sis of Layered Social Networks" [47], where he presents methods for obtaining information characteristics for uncooperative network models. He does this by combining weights of different actor and network characteristics in order to derive the strength of interpersonal relationships. This work, when applied to our actor models can determine the strength of relationships.

6 Conclusion and Future Work

Social network analysis may one day answer the how and why some ideas become successful in reaching a tipping point. Better logics for describing these events, in terms of social interaction, are needed. We have presented the basic first steps to utilizing tensed predicate logic to describe these events, with emphasis on actor roles, within these networks. We have taken semi-formal definitions from Malcolm Gladwell's three primary archetypes: connectors, mavens, salesmen; and we expanded them to include more distinct sub-types. Our primary contribution is a translation of the qualitative defintions of these actor types into formal representations both in tensed predicate logic and into RDF queries that can be directly run on data collected from Twitter.

Previous work has focused on modeling social networks using graph theoretical methods. These approaches rely on measurements taken at a single point in time with no understanding of relationships, such as liaisons, which take place over a period of time. We have shown that by combining logic systems, such as propositional calculus, predicate calculus and Prior's tense logic, that we can create basic definitions for different actor types within a social network.

We have demonstrated the translation of our logical descriptions into RDF Queries using RDFLib in Python. We have tested these implementations on a select day of captured Twitter data. We have explored options for combining both standard graph analysis methods with semantic querying using ontologies specifically designed for social network analysis.

Our approach to describing actor types as tensed predicate logic is important because it allows for direct representation using RDF queries.

7 Acknowledgement

We thank Joshua Taylor, a doctoral candidate in the Computer Science department at Rensselaer Polytechnic Institute (RPI) for his review and validation of our equations [28]. We also thank George Albert, a computer science student at the University of Rochester for his review and equation validation.

References

[1] Gladwell, M. (2000). "The tipping point". Boston: Little, Brown and Company.

[2] Allen, H. T. (1976). "Communication networks - The hidden organizational chart". The Personnel Administrator, 21(6), 31-35.

[3] Arun Phadke, James Thorp. (1978). "Contracts and Influence". Social Netowrks, 1:1-48

[4] Davis, A., et. al. (1941). "Deep South: A social Anthropological Study of Caste and Class". University of Chicago Press. Chicago, Ill.

[5] Freeman, L. (2004) "The Development of Social Network Analysis: A Study in the Sociology of Science". BookSurge, LLC. North Charleston, SC.

[6] Stanley Wasserman, Katherine Faust. (1994). "Social Network Analysis: Methods and Applications", Structural Analysis in the Social Sciences, 25 November 1994

[7] Donald Triner. (2010). "Publicaly Available Social Media Monitoring and Situational Awareness Initiative," Office of Operations Coordination and Planning: Departmetn of Homeland Security, June 22 2010.

[8] Juris, Jeffrey. (2012). "reflections on #Occupy Everywhere: Social media, public space, and emerging logics of aggregation". American Ethnologist. Vol 39, No. 2, pp. 259-279.

[9] Sheedy, Caroline. (2011). "Social Media for Social Change: A Case Study of Social Media Use in the 2011 Egyptian Revolution". Capstone Project.

[10] Stark, Rodney. (1987). "Deviant Places: A Theory of the Ecology of Crime". Criminology, 25: 893910.

[11] Brett Stone-Gross, et al. (2011). "The underground economy of spam: a botmaster's perspective of coordinating large-scale spam campaigns." In Proceedings of the 4th USENIX conference on Large-scale exploits and emergent threats (LEET'11). USENIX Association, Berkeley, CA, USA, 4-4.

[12] Taylor Dewey, et al. (2012). "The Impact of Social Media on Social Unrest in the Arab Spring". Stanford University - Defense Intelligence Agency Final Report.

[13] Christian Sturm and Hossam Amer. (2013). "The effects of (social) media on revolutions: perspectives from egypt and the arab spring". In Proceedings of the 15th international conference on Human-Computer Interaction: users and contexts of use - Volume Part

III (HCI'13), Masaaki Kurosu (Ed.), Vol. Part III. Springer-Verlag, Berlin, Heidelberg, 352-358.

[14] Woods, Richard. (2010). "Privacy is Dead?: Facebook's Mark Zuckerberg says privacy is dead. So why does he want to keeps this picture hidden?". Times Newspapers Ltd.

[15] Statistics Brain. (2013). "'Facebook Statistics". Statistic Brain Research Institute, publishing as Statistic Brain. 6/23/2013. http://www.statisticbrain.com/facebook-statistics/

[16] CBS News. (2012). "Twitter's censorship plan rouses global furor". Associated Press. January 27, 2012

[17] Statistics Brain. (2013). "Twitter Statistics". Statistic Brain Research Institute, publishing as Statistic Brain. 5/7/2013. http://www.statisticbrain.com/twitter-statistics/

[18] Bagley, Nick. (2012). "The Decline of Myspace: Future of Social Media". Dreamgrow Digital. 8/13/2012. http://www.dreamgrow.com/the-decline-of-myspace-future-of-social-media/

[19] alton, Antony, "Temporal Logic", The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), Edward N. Zalta (ed.)

[20] Shea Bennett. "Just How Big Is twitter In 2012 [INFOGRAPHIC]". All Twitter - The Unofficial Twitter Resource, February 2013

[21] Mallon, Shanna. (2012). "50 Facts about Social Media for Business". Straight North, LLC publishing as The Straight North Blog. Downers Grove, IL.

[22] D. Karaiskos, et. al. (2010) "Social network addiction : a new clinical disorder?". European psychiatry : the journal of the Association of European Psychiatrists. volume 25, Page 855. DOI: 10.1016/S0924-9338(10)70846-4)

[23] Helms, R, Ignacio, et al.(2010) "Limitations of Network Analysis for Studying Efficiency and Effectiveness of Knowledge Sharing" Electronic Journal of Knowledge Management Volume 8 Issue 1 (pp53 - 68)

[24] Dhar, Vasant. (2013) "Data Science and Prediction". Communications of the ACM. Vol. 56 No 12, Pages 64-73. 10.1145/2500499

[25] Sullivan, Danny. (2011). "Why Second Chance Tweets MAtter: After 3 Hours, Few Care About Socially Shared Links". Thrid Door Media Inc. Publishing as Search Engine Land.

[26] Travers J., Milgram S. (1969) "An Experimental Study of the Small World Problem," Sociometry, Vol. 32, No. 4. pp. 425-443, doi:10.2307/2786545

[27] Harrist, A. W., Zaia, A. F., Bates, J. E., Dodge, K. A. and Pettit, G. S. (1997). "Subtypes of Social Withdrawal in Early Childhood: Sociometric Status and Social-Cognitive Differences across Four Years". Child Development, 68: 278294. doi: 10.1111/j.1467-8624.1997.tb01940.x

[28] Taylor, J. (2013). "Personal communication". August 12, 2013.

[29] Galton, Antony. (2008). "Temporal Logic". The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.). URL = http://plato.stanford.edu/archives/fall2008/entries/logic-temporal/.

[30] Minker, Jack. (1982). "On indefinite databases and the closed world assumption". Lecture Notes in Computer Science. 6th Conference on Automated Deduction. Springer Berlind Heidelberg. pp. 292-308 doi=10.1007.BFb0000066

[31] Jeremy J. Carroll, Ian Dickinson, Chris Dollin, Dave Reynolds, Andy Seaborne, and Kevin Wilkinson. (2004). "Jena: implementing the semantic web recommendations," In Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters (WWW Alt.'04). ACM, New York, NY, USA, 74-83. DOI=10.1145/1013367.1013381

[32] Claudio Gutierrez, et al. (2005) "Temporal RDF". In Proceedings of the Second European conference on The Semantic Web: research and Applications (ESWC'05), Asuncion Gomez-Perez and Jerome Euzenat (Eds.). Springer-Verlag, Berlin, Heidelberg, 93-107. DOI=10.1007/11431053˙7 http://dx.doi.org/10.1007/11431053˙7

[33] Andrew Page.(2012). "Know Your Meme: Kony 2012". http://www.knowyourmeme.com/memes/events/kony-2012

[34] Goutam Kumar Saha. 2007. "Web ontology language (OWL) and semantic web." Ubiquity 2007, September, Article 1 (September 2007), 1 pages. DOI=10.1145/1295289.1295290 http://doi.acm.org/10.1145/1295289.1295290

[35] John Guare, "Six Degrees of Seperation," A Play, May 1990

[36] Lada Adamic, et al. (2003). "A social network caught in the Web," First monday, 8(6)

[37] David Liben-Nowel, et al. (2005). "Geographic Routing in Social Networks," Proceedings of the National Academy of Sciences (PNAS), 102:11623-1162, 2005

[38] Ravi Kumar, et al. (2006). "Structure and Evolution of Online Social Networks," In the Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mininig (KDD;06), Philadelphia, PA.

[39] Michelle Girvan, Mark Newman. (2002). "Community structure in social and biological networks," Proceedings of the National Academy of Sciences (PNAS), 99(12):7821-7826.

[40] Ceren Budak, et al. (2010). "Where the blogs tip: connectors, mavens, salesmen and translators of the blogosphere". In Proceedings of the First Workshop on Social Media Analytics (SOMA '10). ACM, New York, NY, USA, 106-114. DOI=10.1145/1964858.1964873

[41] Steven Levitt, Stephen J. Dubner. (2005) "Freakonomics: A Rogue Economist Explores the Hidden Side of Everything," New York: Morrow-Harper.

[42] George Kelling, Catherine Coles. (1998). "Fixing Broken Windows: Restoring Order and Reducing Crime in Our Communities," January 20, 1998

[43] Roe v. Wade, 410 U.S. 113 (1973)

[44] Jonah Beger. (2013). "Contagious: Why Things Catch On," Simon and Schuster Publishing, March 5, 2013

[45] R. S. Renfro. (2001). "Modeling and Analysis of Social Networks',' PhD thesis, Air Force Institute of Technology.

[46] C. Clark. (2005). "Modeling and analysis of clandestine networks," Masters thesis, Air Force Institute of Technology.

[47] J. T. Hamill. (2006). "Analysis of Layered Social Networks," PhD thesis, Air Force Institute of Technology.

[48] G. Ereteo , F. Gandon, M. Buffa, O. Corby. (2009) "Semantic Social Network Analysis," Proceedings of the WebSci09. http://journal.webscience.org/141/