Measuring the Dynamic Bi-directional Influence between Content and Social Networks

VU University Amsterdam
De Boelelaan 1081a, 1081 HV, Amsterdam, The Netherlands
{swang,pgroth}@few.vu.nl

Abstract. The Social Semantic Web has begun to provide connections between users within social networks and the content they produce across the whole of the Social Web. Thus, the Social Semantic Web provides a basis to analyze both the communication behavior of users together with the content of their communication. However, there is little research combining the tools to study communication behaviour and communication content, namely, social network analysis and content analysis. Furthermore, there is even less work addressing the longitudinal characteristics of such a combination. This paper presents a general framework for measuring the dynamic bi-directional influence between communication content and social networks. We apply this framework in two use-cases: online forum discussions and conference publications. The results provide a new perspective over the dynamics involving both social networks and communication content.

1 Introduction

Does an informative post on a microblogging service lead to a user gaining followers? If a user is popular in a social network, will their new status updates be widely quoted? If a researcher identifies a new topic one year, does that result in the research having more coauthors the next? As an increasing amount of content is mediated through social networks, these types of questions are of great interest, in particular, to developers, social scientists, and business that aim to understand the link between content generation and social connection. A key aspect to answering these questions is to understand how the relationships between users influence the content of their communication and vice versa.

In this paper, we extend our work in [26] by proposing a general framework for measuring such influence over time. In our approach, we translate both user relationships and content into two corresponding networks: a social network and a content networks. The networks are then characterized using common network properties such as (in-/out-)degree and betweenness centrality. The influence is then measured using a set of multilevel time-series regression models producing what we term an influence network showing how these variables impact each other in time. Additionally, our Influence Framework can integrate other network properties tailored to a given problem domain.

P.F. Patel-Schneider et al. (Eds.): ISWC 2010, Part I, LNCS 6496, pp. 814–829, 2010.

The use of the Influence Framework is facilitated by the emergence of Semantic Web technologies not only to represent relationships between users on the Social Web but also to link to the content those users exchange. For example, the Semantically Interlinked Online Communities (SIOC) ontology is for the representation of the content of discussions but is explicitly intertwined with the Friend of a Friend (FOAF) ontology that is used to represent personal relationships. Because the Semantic Web provides these explicit links, it is easier to obtain the input data sets required by our Influence Framework. Thus, as more Social Web content is made available using Semantic Web standards, the Framework can be used to investigate a wider variety of content and social networks. Later, we show how the Influence Framework can be applied to networks obtained by querying the Semantic Web Dog Food dataset [24] as well as networks extracted from a Dutch political forum. The ability to study the connection between people through their objects was posited as a key benefit to the Social Semantic Web [5]. This work is an example of where these benefits are coming to fruition.

In summary, the contributions of this paper are as follows:

- A general framework for measuring the bi-directional influence between networks of people and the content associated with those people.
- A multilevel time-series regression model for measuring the longitudinal influences between the network properties of content and social networks.
- The generation of influence networks for both Dutch political forums and the World Wide Web conference series, which provide new material for social scientists to investigate these domains.

The rest of this paper is organized as follows. We begin by presenting the Influence Framework and its constituent parts. This is followed by a discussion of the application of the Framework to two use cases: one studying a conference series and the other studying data from a Dutch political forum. Related work is then discussed followed by a conclusion.

2 Influence Framework

The Influence Framework is a three stage framework for measuring the influence between (and within) user relationships and the content they communicate. While such measures of influence are clearly possible to perform on a case-by-case basis, a key realization in this work is that by representing content and user relationships as networks, standard network properties can provide a good initial insight into influence in different domains. We note that influence is a time-dependent notion and thus our framework requires time series data.

The three stages of the framework are:

1. Network Generation
2. Measuring Network Properties
3. Time Series Analysis

We now discuss each of these stages.

2.1 Network Generation

The first stage of the framework is to generate a series of both content and social networks as well as bindings between those networks. The starting point is information about a set of actors who interact over time, *e.g.* , participants in online discussions, scientists who co-author, *etc.* . From these data sets, *a series of social networks* representing the interaction of these actors over time can be produced. Then, a corpus of content related to each actor produced over time is needed *e.g.* , the textual content of online discussions a participant posted, the abstract a scientist wrote, the movies a star acted in, *etc.* . This content corpus should also have the property that pieces of content are somehow similar across a group of actors. Based on some similarity measure between content at each time step, a *series of content networks* can be generated. A key artifact for the framework is documentation of the relationship between actors and the content they produce at each time step. We term these *bindings*.

The network generation stage is perhaps the most domain specific part of the framework as a decision must be made about which content and which sort of user relationship should be represented in the network. Furthermore, many domains have different data formats requiring specialized programs to generate the needed networks. This is where Social Semantic Web technologies are particularly important. By providing common query interfaces and data representations, the extraction of these networks is significantly easier as demonstrated in Section 3.1.

2.2 Measuring Network Properties

Once the content networks and social networks have been produced, the properties of those networks that are of interest need to be defined (as variables) and then measured. The necessary requirement of these properties is that they vary over time. Because the content and social relationships are defined as networks, common network properties can be measured first. For a graph $G = (V, E)$ with a set of vertices $V = \{v_1, \ldots, v_n\}$ and a set of edges $E = \{e_{ij} \mid 1 \leqslant i, j \leqslant n\}$, the common network properties suggested are:

Degree centrality. For a given vertex v_i, its degree centrality is equal to the degree of v_i divided by the maximum possible degree. That is, the degree centrality $C_D(v_i)$ for vertex v_i is:

$$C_D(v_i) = \frac{deg(v_i)}{n-1}$$

In a directed network, two separate measures of degree centrality, namely **in-degree** and **out-degree**, should be measured instead.

Betweenness centrality. The betweenness centrality of a vertex is defined as the fraction of all shortest paths that pass through it over all shortest paths in the network. That is,

$$C_B(v_i) = \sum_{\substack{v_s \neq v_i \neq v_t \in V \\ v_s \neq v_t}} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

where σ_{st} is the number of shortest paths from v_s to v_t (v_s, $v_t \in V$) and $\sigma_{st}(v_i)$ is the number of shortest paths from v_s to v_t that pass through v_i.

Clustering coefficients. Our analysis is at the vertex level, therefore, we measure the local clustering coefficient of a vertex which quantifies how close its neighbors are to being a clique (complete graph). It is measured as the proportion of links between the vertices within its neighbourhood divided by the number of links that could possibly exist between them. Let N_i be the neighbourhood of vertex v_i, *i.e.* , its immediately connected neighbours. For directed graphs, the local clustering coefficient of vertex v_i is given as

$$CC(v_i) = \frac{|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E.$$

While for undirected graphs, it is defined as

$$CC(v_i) = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E.$$

A higher $CC(v_i)$ means the neighbours of v_i are more densely connected.

It is important to note that while these network properties can be measured for every graph, their underlying meaning with respect to the social reality needs to be defined on a per domain basis.

While these measures are a useful start, any network property that varies over time is allowable within the Influence Framework. Later, we show how other more domain specific network properties can be used to gain additional insight into the influence between content and social networks.

The output of this stage is a table mapping each actor to values for each property at each time step.

2.3 Multilevel Time-Series Regression Models

Our Framework aims to model the longitudinal influences between network properties derived from both social and content networks. The output of Stage 2 provides data at successive time steps spaced at uniform time intervals, which form a *time series*. Thus, we need to apply *time series analysis* to extract meaningful statistics of the data in order to better understand the underlying forces and structures that produced the observed data. By fitting to a time series model, we can proceed to forecasting and predicting the forthcoming data [30]. When modeling variations in the level of a process, one of the typical methods is to use the *autoregressive* (AR) models.

Let $\mathbf{X}$ be a time series: $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$, where $\mathbf{x}^{(t)}$ is the data observation at time t. Here, $\mathbf{x}^{(t)}$ is a vector, *i.e.* $\mathbf{x}^{(t)} = (x_1^{(t)}, x_2^{(t)}, \ldots, x_m^{(t)})^T$, where m is the total number of variables we are modelling and each $x_i^{(t)}$, $i = 1, \ldots, m$, is a variable we are interested in, such as the betweenness and degree centrality of a node in the social network or the centrality values of certain political or scientific topics. The $AR(p)$ model is defined as

$$x^{(t)} = a + \sum_{j=1}^{p} b_j \, x^{(t-j)} + \varepsilon^{(t)}, \tag{1}$$

where $b_1, \ldots, b_p$ are the parameters of the model, a is a constant and $\varepsilon^{(t)}$ is the noise with Gaussian distribution. In this paper, we opt for a simple model for each variable x_i independently, which only includes the values from the last time-point as independent variables, *i.e.* , an AR(1)-process:

$$x_i^{(t)} = a_i + b_{1i} \, x_1^{(t-1)} + \cdots + b_{mi} \, x_m^{(t-1)} + \varepsilon_i^{(t)}, \tag{2}$$

where $\varepsilon_i^{(t)}$ is Gaussian noise with zero mean and variance σ_ε^2.

In these models, each variable $x_i^{(t)}$ at time t is modelled as a linear combination of the predictor variables at time $t-1$, each weighted by a coefficient that quantifies how variation in the predictor variable at time $t-1$ is related to the variation of the predicted variable at time t. Such coefficients or *effects* can tell us the influence among different variables over time.

Generally, the above mentioned variables are referred to in statistics as *units of analysis*. In social reality, these variables are often from different levels, which are frequently hierarchically nested. For example, when studying the research achievements, attributes of individual researchers, research groups, faculties and the universities as a whole can all be important units of measures. This stage applies the above introduced regressive model to study the influence between variables, and the resulting coefficients are also called *fixed* effects. However, there exist variations among different actors, *i.e.* , *random* effects (actor-level errors). Therefore, such single-level statistical methods are no longer appropriate to study these so-called *complex data sets* [31]. We thus need to apply *multilevel analysis* to examine both fixed and random effects of variables measured at different levels [13,31].

Formally, we define $\mathbf{x}_p^{(t)} = (x_{1,p}^{(t)}, \ldots, x_{m,p}^{(t)})^T$, a vector containing the variables for actor p at time t. We can then rewrite equation (2) as

$$x_{i,p}^{(t)} = a_i + \mathbf{b}_i^T \, \mathbf{x}_p^{(t-1)} + \varepsilon_i^{(t)} + \mathbf{c}_{i,p}^T \, \mathbf{x}_p^{(t-1)} + \varepsilon_{i,p}^{(t)}, \tag{3}$$

where $\mathbf{b}_i = (b_{i1}, \ldots, b_{im})^T$ and $\mathbf{c}_i = (c_{i1}, \ldots, c_{im})^T$ are the fixed-effect coefficients and random-effects coefficients respectively.

In order to compare the resulting fixed effects to each other, all variables in the random effects regression equations need to be linearly transformed into standardised values, *i.e.* , subtraction of their mean, division by their standard deviation. In this way, the fixed effects can be interpreted as the effect of one standard deviation of change in the independent variable on the number of standard deviations change in the dependent variable.

The output of this stage is the set of statistics generated in fitting the regression models as well as a diagram, called an *influence network*, that shows the statistically significant effects between variables.

3 Use Cases

We now present two use cases applying the Influence Framework. First, a simple use case based on existing Semantic Web data is discussed. It analyses the influence between co-authorship and the topics addressed at a conference. The second use case looks at the influence of social status of forum participants and their focus on particular political parties. This use case is then extended to consider newly defined variables to answer specific questions of the domain.

3.1 Influence between Co-authors of Academic Papers and the Topics They Address

Data Collection. The World Wide Web Conference is the preeminent conference on Web Technologies covering both advances in academia and industry. We obtained a corpus of metadata about this conference from the Semantic Web Dogfood repository [24]. The metadata covers the conference program including paper metadata (*e.g.* , authors, paper titles, keywords, *etc.*) and organization metadata (*e.g.* , program committee members, collocated workshops, *etc.*). Importantly for use with the Influence Framework, the metadata spans four years of the conference from 2007 to 2010 using generally the same schema. The data was downloaded in bulk and loaded into separate RDF stores for each year.

Generating Social Networks. We chose the co-author network as the social network of interest. For every year, we retrieved the co-author pairs for each article using the SPARQL query shown in Figure 1. From these results, we built a weighted undirected graph for each year where nodes are authors, edges are shared authorship of an article and the weights on edges are the number of co-authorships between the two linked authors. For wider coverage, we did not distinguish between paper types that is a workshop, main track, or poster paper are all considered equal for the purposes of co-authorship.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX swrc: <http://swrc.ontoware.org/ontology#>

SELECT ?author ?coauthor ?article WHERE {
      ?article swrc:author ?author.
      ?article swrc:author ?coauthor
}
```

Fig. 1. Query to extract co-author pairs

For each year, we measured the degree and betweenness centrality of each of the authors. The degree centrality represents how activity the author is in coauthoring with others. Clustering coefficient provides a measure for how closely knit a group is. In this case, it provides a measure of whether authors write with the same set of other authors. For example, one can imagine that the authors from the same department may form a cluster within the co-author network.

Generating Content Networks. Here, we are interested in the topics under discussion at the conference in each year. To obtain those topics, we use author assigned keywords as proxies for those topics. This is common practice within the bibliometrics community [3]. Similar to the co-author network, we retrieved the keywords for each article in the conference via a SPARQL query. To improve overlap between keywords assigned by different authors, keywords containing more than one word were split into separate words and then stemmed. Stemming allows keywords such as ontologies and ontology to be treated the same. Based on the stemmed keywords, a weighted undirected graph is built, where a node is a keyword and an edge is the co-occurrence between two keywords in the set of keywords for an article. Edges are weighted by the number of co-occurrences. A graph is produced for each year.

As prescribed by the Influence Framework, we then compute several common network metrics. Again, the degree provides information about the popularity of a given topic. The betweenness centrality provides information about whether a keyword is a bridge between two other keywords (*i.e.* topics).

Binding Social Content Networks. We bind the two networks together via the papers within the conference. Thus, we know which author discusses a topic and what topics are associated with particular authors via their connection to papers.

Influence Network. For this use case, we use five network measures.

- Three social network properties: degree centrality, betweenness centrality, and clustering coefficient.
- Two content-wise properties: degree centrality, betweenness centrality.

The units of analysis are all year × participant combinations. The multilevel time-series regression models are then constructed to to study the influence network between topics of a conference and the co-authorship of papers.

Figure 2 shows the resulting influence network. This network only shows effects which are statistically significant. Note, when reading such an influence network, the edges are directional in time. For example, in Figure 2, the edge between degree in the content network and clustering coefficient in the social network, should be read as the degree at some time t has large negative effect on the clustering coefficient in time $t + 1$.

The network suggests a number of avenues for investigation. First, there is strong negative effect between the degree centrality of a topic (*i.e.* , keyword) on itself, which suggests that a popular topic one year is likely to be less popular the next. Degree centrality of a topic also has strong negative effects on the degree centrality and clustering coefficient for an author. One interpretation of this

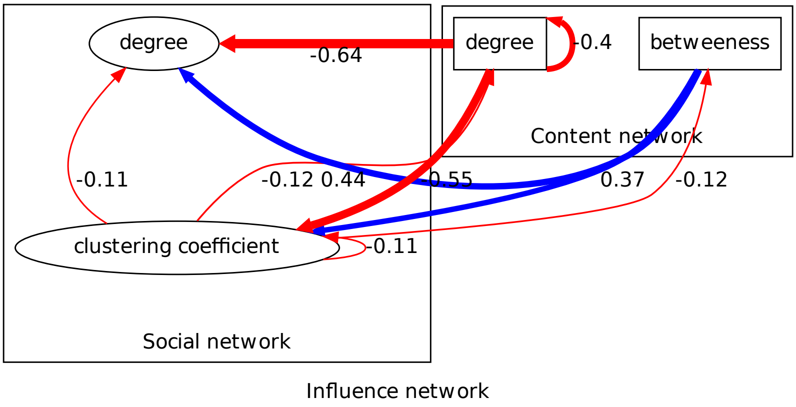

Fig. 2. Influence network for WWW conference

result is that after a burst of collaboration on a hot topic, the topic becomes less exciting and the collaboration between authors around it dies down. There are strong positive effects of the betweenness centrality of a topic and the subsequent degree centrality and clustering coefficient of an author. A possible explanation for these effects is that if a topic bridges the gap between other topics in one conference year, it is likely to become the focus for new collaborations between authors concentrating on these normally separate topics. Such new collaborations would then come to the foreground in the next conference year.

3.2 Influence between Social Status of Online Forum Participants and Their Political Attention

Data Collection. Our data is collected from the biggest and one of the oldest Dutch forums, NL.politiek, which is entirely devoted to politics. This forum has more than 40,000 participants. Our dataset contains all the postings from October 2003 to December 2008, in total more than 1.1 million postings.

Generating Social Networks. All postings were divided into weekly subsets. In each subset, all postings were grouped by their threads and ranked based on their time stamps. Each thread corresponded to a mini discussion network, where the participants reacted to others by replying to their postings. Formally, a mini discussion network (*i.e.* , a thread) is a graph $G = (V, E)$, where V is a set of participants in this thread, and E the weighted and directed connections between the participants. There is a directed link (v_i, v_j) if participant v_i replied at least once to one of the postings of participant v_j. The frequency of the occurrence of such replying action was considered as the weight of the link, $w(v_i, v_j)$. Note, online participants often post more than once in the same thread, replying to previous postings which may include their own. Therefore, such networks can be reflexive.

We then aggregated all the mini discussion networks within one week into a bigger network, producing a series of 259 weekly social networks where 21,127 participants are involved. We note that the extraction of these networks would have been greatly simplified if they had been represented using SIOC, for example.

For each week, we measured the in-/out-degree and betweenness centrality of all participants. In this setting, the in-degree centrality of a participant indicates the degree of *popularity* he has in the online community. The out-degree centrality indicates how active one participant is. The betweenness centrality is an indicator of the *mediating/brokerage* role of a participant. A high betweenness centrality suggests that the participant connects separate communities. The brokerage role of the persons with a higher betweenness centrality is the key to understand the structural hole theory of organisational communication [9].

Generating Content Networks. In this use case, we are interested in the attention to the political parties that online participants have when they discuss in the forum. We thus extract the co-occurrence of parties as the content network. Since co-occurrence is symmetric, the content networks are therefore undirected. In the content network, the vertices are 19 Dutch parties, *i.e.* , $V = \{p_1, \ldots, p_{19}\}$. At the weekly basis, for each party p_i, we gathered a set of postings where the party was mentioned,[1] noted as S_{p_i}, $i = 1, \ldots, 19$. The weight of the edge (p_i, p_j) is calculated as the Jaccard similarity coefficient between two sets S_{p_i} and S_{p_j}, that is,

$$w(p_i, p_j) = \frac{S_{p_i} \cap S_{p_j}}{S_{p_i} \cup S_{p_j}}$$

In this way, we also extracted 259 weekly content networks. We then measured the betweenness and degree centrality of each party in each week. These centrality can tell us how one party's popularity and breakage role evolves over time. When a party has a higher degree centrality, then this party is more often mentioned while other parties are being discussed, *i.e.* , this party is more relevant or important. A party with a higher betweenness centrality is more often mentioned as a reference while more than two parties are mentioned.

Binding Social and Content Networks. We bind two networks based on *who talked when, about what.* For each participant, we counted how many times he talked about one or more of the 19 Dutch parties in a particular week, noted as $\{O_{p_1}, \ldots, O_{p_{19}}\}$. Then the degree centrality of this participant in terms of his discussion content is calculated as

$$cdc = \sum_{1}^{19} O_{p_1} \times dc(p_i)$$

[1] This is done through the AmCAT tool (`http://content-analysis.org/`) which uses a dictionary of keywords to signify an occurrence of a party when one of its keywords is used in the posting.

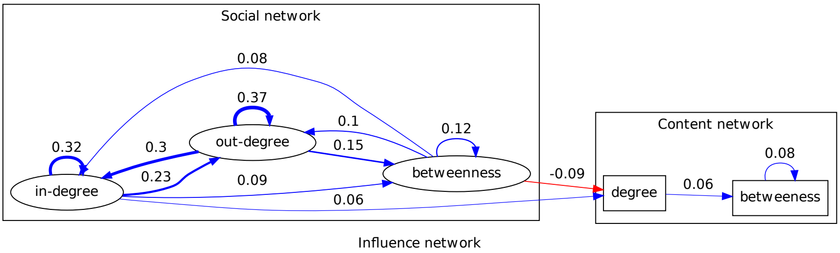

Fig. 3. Standard influence network

where $dc(p_i)$ is the degree centrality of Party p_i in the extracted content network of this week. The betweenness centrality in terms of the content, cbc, is calculated similarly.

Influence Network. Similar to the conference case, we have five standard network variables to model:

- Three social network properties: sbc (betweenness centrality), $sidc$ (in-degree centrality) and $sodc$ (out-degree centrality)
- Two content-wise properties: cbc (betweenness centrality) and cdc (degree centrality)

The units of analysis are all the week × participant combinations. We built the multilevel time-series regression models as introduced in Section 2.3 to study the influence network among political attention and social status in the online community.

There are 1762 participants have posted more than 10 postings during the whole period of time. Therefore, the Figure 3 is based on 433,453 observations from these 1762 participants. The value on the links are the fixed effects, with the critical value $p < 0.05$.

Not surprisingly, the in and out degree centrality have positive effect upon each other and to themselves. When a participant is more active, they are also likely to be more popular and more active in the social network, and vice versa. Also, the two degree centralities and the betweenness centrality have positive effects on each other with the similar strength. Once a participant gains a relatively strong brokerage role, they are more likely to maintain this role, by continuing to react to others, which consequently causes more people to reply to them. Looking at the effects between social network and the content network, the in-degree centrality (*i.e.* , the popularity of a participant) has a positive effect on the degree centrality of the content. This suggests that when a popular participant talks about certain parties, these parties are likely to become popular in the next week. When a participant becomes a broker, they tend to communicate with different opinion-holders, therefore they discuss more parties instead of only popular ones. This might be the reason for the negative effect from the social betweenness centrality to the content degree centrality. However, this may also be because of the correlations between these fixed effects, which needs to be further investigated.

3.3 Influence between User-Defined Content Variables with Social Network Properties

Content networks can be extracted in a manner that is more suitable to specific problems within a domain. Communication scientists are interested in not only the attention that the online forum participants pay to the political parties, but also the degree to which they follow the agenda of the mass media. Online discussions are expected to be more emotional and more aggressive (negativity, hatred, disgust, in short *flaming*) as compared to the news from the mass media [25]. It is natural for the communication scientists to ask to which degree emotions and aggression are expressions of autonomous or even anarchistic of online participants, to which degree they are caused by the news content in the mass media, as the classic theory of agenda setting would suggest, and to which degree they reflect depersonalised, scale-free properties of the social network of online participants that can be predicted from the previous state of their social networks.

Data Collection. We further collected newspaper articles from five biggest Dutch national newspapers. The selected national newspapers (Telegraaf, NRC Handelsblad, Algemeen Dagblad, de Volkskrant and Trouw) represent mainstream politics in the Netherlands. These newspapers reach one third of the Dutch population (official figures in 2008, http://www.cebuco.nl). The newspaper articles were retrieved from the LexisNexis archive,[2] each of which mentioned at least one political actor (*e.g.* , Dutch politicians or parties). We took a random subset of newspaper articles published between 2006 and 2008. Therefore, we also take 157 weekly social networks in these three years into the analysis.

Extracting Content Variables. In this paper, we focus on two aspects related to the forum content. The first aspect is related to the agenda setting [29,33]. The agenda setting hypothesis maintains that the participants in the online environment will take over the issue agenda from the mass media in a top-down fashion. An alternative hypothesis is that the mass media nowadays take over the topics raised in online discussion forums, in order to express and disseminate the opinions of their audience to decision-makers in business and politics, or in order to keep their audience in competitive media markets. Here, we are interested in whether the social status of the participants is influenced by the extent to which they follow mass media. Therefore, we use a list of political issues and measure weekly the attention to these issues (the frequencies of occurrence of these issues) in the newspaper articles and online discussions, respectively. Then a correlation is calculated between these two lists of the attention, which gives the first content

[2] http://www.lexisnexis.com/

variable *NewspaperContagion*. A higher *NewspaperContagion* indicates that the participant more strongly follows the agenda of the newspapers.

Another interesting aspect is the above-mentioned emotion expressed in the forum discussions. We would like to check whether the amount of emotion expressed in the online discussion influences the social status of the participants and his willingness to following the mass media. Starting from Brouwers thesaurus for Dutch [7], a list of keywords was developed for each emotion. Similar to measuring the attention to political parties, the frequencies of occurrence of these keywords were also measured. We separated the emotion of disgust and hate as a separate variable as they are the major emotions the communication scientists are studying [25]. Therefore, we have two other content variables: *DisgustHate* and *OtherEmotions*.

Influence Network. The five variables we investigate are

- Two network properties: *IPopularity* (=indegree centrality) and *CBetweenness* (betweenness centrality)
- Three communication contents: *DisgustHate*, *OtherEmotions* and *NewspaperContagion*.

Similar multilevel time-series regression models were built to study the influence between these variables. The resulting influence network is shown in Figure 4, based on 171,756 observations from 1101 participants.[3]

Similar to Figure 3, the betweenness centrality and popularity (in-degree centrality) have strong positive effects on themselves and each other. As we can see, a popular member or a brokerage member has a strong tendency to express emotions in their postings, and such emotional expressions also increase their social status. Especially the social popularity and the usage of the language of disgust and hate have impressively strong effects on each other. It may be the case that online participants who feel that they are in the centre of the debate, as measured by a high popularity and betweenness centrality, feel unhindered or even obliged to use rather crude words to maintain their position.

In our dataset, there seems no significant effect from the degree one follows the mass media to the aptitude for flaming and blaming, which is suggested on the basis of the classic agenda setting theory [21].

The decision of following newspaper agenda is influenced by the previous popularity in the community and also the expression of disgust and hate. This finding corresponds with earlier findings that especially citizens who are preoccupied with negativity will like the current type of news, especially men like negative news [12]. A new finding is that a high popularity in an online discussion forum also contributes to taking over agenda cues from the mass media. Apparently popular participants feel inclined to follow the news and to take over the news agenda. This corresponds with the old idea that opinion leaders in a group tend to follow the mass media closely.

[3] Again, only these 1101 participants have more than 9 postings within these three years.

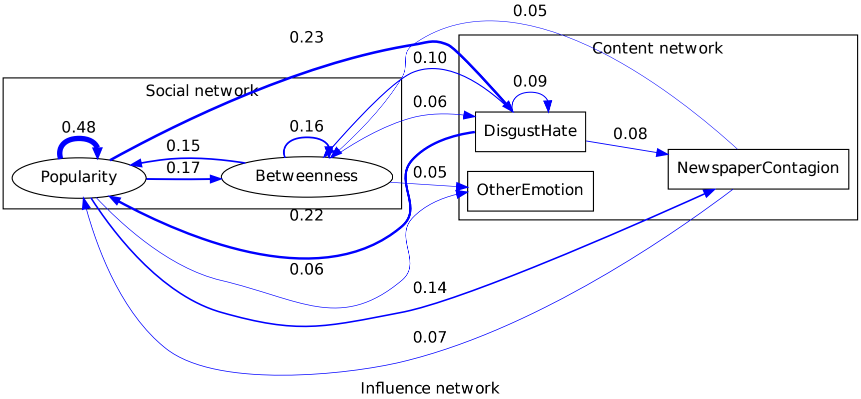

Fig. 4. Regression model of user-defined variables

4 Related Work

Social network analysis (SNA) has recently become a popular topic of study in organisation studies, communication studies, information science, *etc.* It views social relationships in terms of network theory consisting of nodes and ties. Using graph algorithms, SNA characterises the structure of social networks, strategic positions in these networks, specific sub-networks and decompositions of people and activities [28]. SNA has been applied not only to Web 2.0 platforms such as Facebook [1] and wikis [32], but also directly to the whole Web, the blogsphere, ontologies and the Semantic Web [16,17,15]. Recently, Semantic Web techniques have been adopted to facilitate standard SNA procedures [23,20,10].

On the other hand, content analysis is a research tool which has been used since the mid-1950's to determine the presence of certain words or concepts within texts [4,18]. By quantifying and analysing the presence, meanings and relations of such words and concepts, social scientists can make inferences about the content of the texts. As it is applicable to any piece of writing or recorded communication, it has been widely used in many fields, such as media studies, literature, sociology and political science [14,8,34]. Recently, many efforts have been focused on automated content analysis, such as [2], which to a large degree improves the access to large corpora.

These two classes of analysis have been investigated and applied in a rather parallel style. Only until recently, social scientists started to combine social network analysis and content analysis, such as the *discourse network analysis* in [19], and the work in [27]. This paper is the first to combine these two kinds of analysis in the Semantic Web context.

Another focus of our paper is on the longitudinal analysis over content and social networks. Recognised as a *Holy Grail* for network researchers, there has been a large degree of focus on the analysis of social networks over time [22]. However, there has not been much work with respect to the longitudinal analysis on the combination of social and content networks. The closest work is that of Gloor et al, who use network analysis over social networks and corresponding content to identify trends , however, they concentrate on a time dependent betweenness measure and do not provide a general framework for a variety of network properties [11]. Our previous work in [26] is extended in this paper by providing a general framework which is suitable for the analysis of the longitudinal influence between social networks and communication content in the Semantic Web context.

5 Conclusion

In this paper, we presented a general framework for analyzing the dynamic bi-directional influence between social relationships and the content produced with respect to those relationships. The Influence Framework leverages a key insight that by representing both social relationships and content as networks, common network properties can be used to bootstrap the analysis of influence. Based on these properties, the framework applies a time-series regression model to generate influence network diagrams representing the statistically significant effects of these properties. We applied our framework to two domains, dutch politics and a conference series, resulting in interesting conclusions about the influence of media on political forum participants and the impact of topics on academic collaboration. The data was acquired from both a web crawl and a Semantic Web source, we note that the acquisition of networks was easier using the Semantic Web data source. *To the best of our knowledge, this is the first work that combines longitudinal social network analysis and content analysis in the context of the Semantic Web.* In future work, we aim to expand the integration with Semantic Web data sources by providing reusable modules for widely used ontologies such as SIOC. Additionally, we aim to provide a service allowing others to more easily apply this framework to their own data sources.

By linking across both content and social networks, the Social Semantic Web is providing a new data source for understanding the relationship between users and the content that they produce [6]. The framework described in this paper provides a new tool for analyzing these relationships from a longitudinal perspective.

References

1. Ackland, R.: Social network services as data sources and platforms for e-researching social networks. Social Science Computer Review 27, 481–492 (2009)
2. van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N.: Parsing, semantic networks, and political authority using syntactic analysis to extract semantic relations from dutch newspaper articles. Political Analysis 16(4), 428–446 (2008)
3. Becker, H.A., Sanders, K.: Innovations in meta-analysis and social impact analysis relevant for tech mining. Technological Forecasting and Social Change 73(8), 966–980 (2006); tech Mining: Exploiting Science and Technology Information Resources
4. Berelson, B.: Content Analysis in Communication Research. Free Press, New York (1952)

5. Bojars, U., Breslin, J.G., Peristeras, V., Tummarello, G., Decker, S.: Interlinking the social web with semantics. IEEE Intelligent Systems 23, 29–40 (2008)
6. Bojrs, U., Breslin, J.G., Finn, A., Decker, S.: Using the semantic web for linking and reusing data across web 2.0 communities. Web Semant. 6(1), 21–28 (2008)
7. Brouwers, L.: Het juiste woord. Standaard betekeniswoordenboek der Nederlandse taal, 7de druk, bewerkt door F. Clacs. Antwerpen: Standaard Uitgeverij (1989)
8. Budge, I., Klingemann, H.D., Volkens, A., Bara, J., Tanenbaum, E.: Mapping Policy Preferences. In: Estimates for Parties, Electors and Governments 1945-1998. Oxford University Press, Oxford (2001)
9. Burt, R.S.: Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge (1992)
10. Ereteo, G., Buffa, M., Gandon, F., Corby, O.: Analysis of a real online social network using semantic web frameworks. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 180–195. Springer, Heidelberg (2009)
11. Gloor, P.A., Krauss, J., Nann, S., Fischbach, K., Schoder, D.: Web science 2.0: Identifying trends through semantic social network analysis. In: CSE (4), pp. 215–222. IEEE Computer Society, Los Alamitos (2009)
12. Grabe, M., Kamhawi, R.: Hard wired for negative news? gender differences in processing broadcast news. Communication Research 33(5), 346–369 (2006)
13. Hayes, A.F.: A Primer on Multilevel Modeling. Human Communication Research 4, 385–410 (2006)
14. Holsti, O.R.: Content Analysis for the Social Sciences and Humanities. Addison-Wesley, Reading (1969)
15. Hoser, B., Hotho, A., Jäschke, R., Schmitz, C., Stumme, G.: Semantic network analysis of ontologies. In: Sure, Y., Domingue, J. (eds.) ESWC 2006. LNCS, vol. 4011, pp. 514–529. Springer, Heidelberg (2006)
16. Jamali, M., Abolhassani, H.: Different aspects of social network analysis. In: IEEE/WIC/ACM International Conference on Web Intelligence, Hong Kong, pp. 66–72 (2006)
17. Kim, H.M., Biehl, M., Buzacott, J.A.: M-ci2: Modelling cyber interdependencies between critical infrastructures. In: Proceedings of 3rd IEEE International Conference on Industrial Informatics, pp. 644–648 (2005)
18. Krippendorff, D.K.H.: Content Analysis: An Introduction to Its Methodology. Sage Publications, Inc., Thousand Oaks (2003)
19. Leifeld, P., Haunss, S.: A comparison between political claims analysis and discourse network analysis: The case of software patents in the european union. In: MPI Collective Goods Preprint. 2010/21 (May 2010)
20. Martin, M.S., Gutierrez, C.: Representing, querying and transforming social networks with rdf/sparql. In: Aroyo, L., Traverso, P., Ciravegna, F. (eds.) Semantic Web: Research and Applications, pp. 293–307 (2009)
21. McCombs, M., Shaw, D.: The agenda-setting function of mass media. Public Opinion Quarterly (1972)
22. McCulloh, I., Carley, K.: Longitudianl dynamic network analysis, using the over time viewer feature in ora. Tech. rep., Institute for Software Research, School of Computer Science, Carnegie Mellon University (2009)
23. Mika, P.: Flink: Semantic web technology for the extraction and analysis of social networks. Journal of Web Semantics 3, 211–223 (2005)

24. Möller, K., Heath, T., Handschuh, S., Domingue, J.: Recipes for semantic web dog food: the ESWC and ISWC metadata projects. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 802–815. Springer, Heidelberg (2007)
25. Oegema, D., Kleinnijenhuis, J., Anderson, K., Van Hoof, A.: Flaming and blaming: The influence of mass media content on interactions in on-line discussions. In: Konijn, E., Tanis, M., Utz, S. (eds.) Mediated Interpersonal Communication. Erlbaum, Mahwah (2008)
26. Oegema, D., Wang, S., Kleinnijenhuis, J.: Dynamics of online discussions about politics: a function of structural network properties, mass media attention or emotional utterances? In: Proceedings of the WebSci10: Extending the Frontiers of Society On-Line. US, Raleigh (April 2010)
27. Oliver, A.L., Montgomery, K.: Using field-configuring events for sense-making: A cognitive network approach. Journal of Management Stuidies 45, 1147–1167 (2008)
28. Scott, J.: Social Network Analysis: A Handbook, 2nd edn. Sage, Newberry Park (2000)
29. Severin, W., Tankard, J.: Communication theories. Pearson, New York (2010)
30. Shumway, R.H.: Applied Statistical Time Series Analysis. Prentice Hall Series in Statistics. Prentice Hall, Englewood Cliffs (1988)
31. Snijders, T.A.B., Bosker, R.J.: Multilevel analysis: an introduction to basic and advanced multilevel modeling. Sage, Thousand Oaks (1999)
32. Tomasev, N., Mladenic, D.: Semantic web wiki: Social network analysis of page editing. In: LuzarStiffler, V., Jarec, I., Bekic, Z. (eds.) Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces, pp. 505–510 (2009)
33. Walgrave, S., Van Aelst, P.: The contingency of the mass medias political agenda setting power: Toward a preliminary theory. Journal of Communication 56, 88–190 (2006)
34. Wimmer, R.D., Dominick, J.R.: Mass Media Research: An Introduction, 8th edn. Wadsworth, Belmont (2005)