Understanding co-evolution of social and content networks on Twitter

Philipp Singer
Knowledge Management Institute
Graz University of Technology
Graz, Austria
philipp.singer@tugraz.at

Claudia Wagner
DIGITAL Intelligent Information Systems
JOANNEUM RESEARCH
Graz, Austria
claudia.wagner@joanneum.at

Markus Strohmaier
Knowledge Management Institute and Know-Center
Graz University of Technology
Graz, Austria
markus.strohmaier@tugraz.at

ABSTRACT

Social media has become an integral part of today's web and allows users to share content and socialize. Understanding the factors that influence how users evolve over time - for example how their social network and their contents co-evolve - is an issue of both theoretical and practical relevance. This paper sets out to study the temporal co-evolution of content and social networks on Twitter and bi-directional influences between them by using multilevel time series regression models. Our findings suggest that on Twitter social networks have a strong influence on content networks over time, and that social network properties, such as users' number of followers, strongly influence how active and informative users are. While our investigations are limited to one small datasets obtained from Twitter, our analysis opens up a path towards more systematic studies of network co-evolution on platforms such as Twitter or Facebook. Our results are relevant for researchers and social media hosts interested in understanding how content-related and social activities of social media users evolve over time and which factors impact their co-evolution.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks; J.4 [**Computer Applications**]: Social and behavioral sciences—*Sociology*

General Terms

Experimentation, Human Factors, Measurement

Keywords

Microblog, Twitter, Influence Patterns, Semantic Analysis, Time Series

1. INTRODUCTION

Social media applications such as blogs, message boards or microblogs allow users to share content and socialize. Hosting such social media applications can however be a costly, and social media hosts need to ensure that their users remain active and their platform remains popular. Monitoring and analyzing behavior of social media users and their social and content co-evolution over time can provide valuable information on the factors which impact the activity and popularity of such social media applications. Activity and popularity are often measured by the growth of content produced by users and/or the growth of its social network. However, as a research community we know little about the factors that impact the activity and popularity of social media applications and we know even less about how users' content-related activities (e.g., their tweeting, retweeting or hashtagging behavior) influence their social activities (i.e., their following behavior) and vice versa.

This paper sets out to explore factors that impact the co-evolution of users' content-related and social activities based on a dataset consisting of randomly chosen users taken from Twitter's public timeline by using a multilevel time series regression model. Unlike previous research, we focus on measuring dynamic bi-directional influence between these networks in order to identify which content-related factors impact the evolution of social networks and vice versa. This analysis enables us to tackle questions such as "Does growth of a user's followers increase the number of links or hashtags they use per tweet?" or "Does an increase in users' popularity imply that their tweets will be retweeted more often on average?".

Our results reveal interesting insights into influence patterns in content networks, social networks and between them. Our observations and implications are relevant for researchers interested in social network analysis, text mining and behavioral user studies, as well as for social media hosts who need to understand the factors that influence the evolution of users' content-related and social activities on their platforms.

2. METHODOLOGY

Since we aim to gain insights into the temporal evolution of content networks and social networks, we apply *time series modeling* [5] based on the work by Wang and Groth [9] who provide a framework to measure the bi-directional influ-

ence between social and content network properties. In this work we apply an *autoregressive model* in order to model our time series data. An autoregressive model is a model that goes back p time units in the regression and has the ability to make predictions. This model can be defined as $AR(p)$, where the parameter p determines the order of the model. An autoregressive model aims to estimate an observation as a weighted sum of previous observations, which is the number of the parameter p. In this work we apply a simple model, which calculates each variable independently and further only includes values from the last time unit. The calculated coefficients of the model can determine the influences between variables over time.

In regression analysis variables often stem from different levels. So called *multilevel regression models* are an appropriate way to model such data. Hence, the measurement occasion is the basic unit which is nested under an individual, the cluster unit. In our datasets we have such a hierarchical nested structure. For each day different properties are measured repeatedly, but all of these values belong to different individuals in our study. If we would apply a simple autoregressive model to our data we would ignore the difference between each user and would just calculate the so-called *fixed effects*, because we can not assume that all cluster-specific influences are included as covariates in the analysis [7]. The advantage of such multilevel regression models is now that they add *random effects* to the fixed effects to also consider variations among our individuals. Since we measure different properties repeatedly for different days and different individuals in our study, our dataset has a hierarchical nested structure. Therefore, we utilize a *multilevel autoregressive regression model* which is defined as follows:

$$x_{i,p}^{(t)} = a_i^T x_p^{(t-1)} + \epsilon_i^{(t)} + b_{i,p}^T x_p^{(t-1)} + \epsilon_{i,p}^{(t)} \quad (1)$$

In this equation $x_p^{(t)} = (x_{i,p}^{(t)}, ..., x_{m,p}^{(t)})^T$ represents a vector, which contains the variables for an individual p at time t. Furthermore, $a_i = (a_{i,1}, ..., a_{im})^T$ represents the fixed effect coefficients and $b_i = (b_{i,1}, ..., b_{im})^T$ represents the random effect coefficients. It is assumed that $\epsilon_i^{(t)}$ and $\epsilon_{i,p}^{(t)}$ is the noise with Gaussian distribution for the fixed and random effects respectively. It has zero mean and variance σ_ϵ^2 [6] [3]. To compare the fixed effects to each other, the variables in the random effect regression equations need to be linearly transformed to represent standardized values. How this is done and how the model is finally applied to our data is described in section 4.

3. DATASET

We chose Twitter as a platform for studying the co-evolution of communication content and social networks, since it is a popular micro-blogging service. We explore one *random dataset* in this work, which was crawled within a time period of 30 days. This random dataset consists of random users from the public timeline who do not have anything special in common.

To generate the random dataset, we randomly chose 1500 users from the public Twitter timeline who we used as *seed users*. We used the public timeline method from the Twitter API to sample users rather than using random user IDs since the timeline method is biased towards active Twitter users. To ensure that our random sample of seed users consists of active, English-speaking Twitter users, we further only kept users who mainly tweet in English, have at least 80 followers, 40 followees and 200 tweets. We also had to remove users from our dataset who deleted or protected their account during the 30 days of crawling. Hence, we ended up having 1.188 seed users for whom we were able to crawl their social network (i.e., their followers and followees) and their tweets and retweets. To identify retweets we used the flag provided by the official Twitter API and to extract URLs we used a regular expression. During a 30 day time period (from 15.03.2011 to 14.04.2011) we polled the data daily at about the same time.

4. EXPERIMENTAL SETUP

The goal of our experiments is to study the co-evolution of social and content networks of Twitter users and influence patterns between them. In order to achieve this we firstly created a social and content network for each specific time point.

Social network: The social network is a one-mode directed network, where each vertex represents a user and the edges between these vertices represent the directed follow-relations between two users at a certain point in time. The constructed social network of seed users only reflects a sub-part of a greater network. Therefore it makes no sense to calculate and analyze specific network properties such as betweenness centrality or clustering coefficient, because these properties depend on the whole network and we only have data available for a certain sub-network.

Content network: The content network at each point in time is a two-mode network, which connects users and tweets via authoring-relations. From these user-tweet networks one can extract specific tweet features, such as hashtags, links or retweet information, and build, for example, a user-hashtag network. It would also be possible to create further types of content networks, such as hashtag co-occurrence networks (see [8] for further types), but we leave the investigation of such network types open for future research.

Overall, the social networks capture the social following relations between users, whereas our content networks account the tweets users publish. Finally, we can connect both networks via their user vertexes, since we know which user in the social network corresponds to which user in the content network and vice versa.

A further step towards our final results is the normalizing of our available data. This is done by subtracting the time-overall mean and dividing the result by the time-overall standard deviation [2]. The fixed effects can now be analyzed as the effect of one standard deviation of change in the independent variable on the number of standard deviations change in the dependent variable [9].

Based on the prepared data, the final model described in section 2 can be applied to identify a potential influences between social and content network properties over time. Table 1 describes each social and content network property

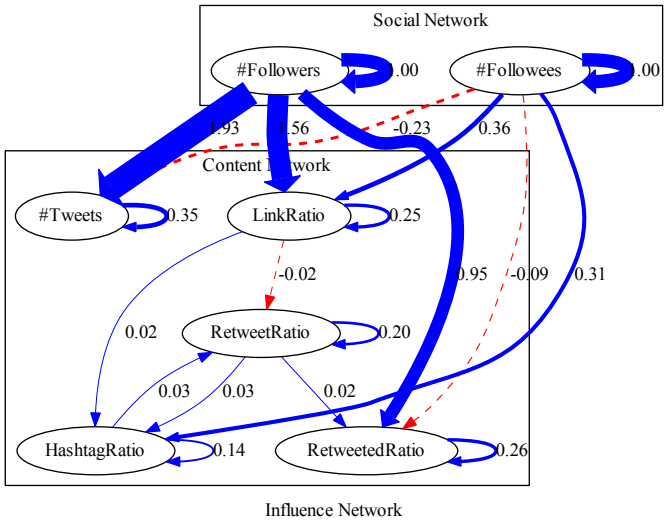

Figure 1: Influence network between the content and social network of a randomly chosen set of Twitter users. An arrow between two properties indicates that the value of one property at time t has a positive or negative effect on the value of the other property at time $t+1$. Red dashed arrows represent negative effects and blue solid arrows represent positive effects. The thickness of the lines indicates the weight of the influence relations. Only statistically significant influences are illustrated.

used throughout our experiment. The properties are calculated for a corresponding social or content network at each time point t of the random Twitter dataset. The dependent variable of the model is always a property at time t and the independent variable are all properties at time $t-1$ including the dependent variable at that time. Including the dependent variable in that step allows us to detect if a variable's previous value influences it's future value. Finally, the resulting statistical significant coefficients show a relationship between an independent variable at time $t-1$ and a dependent variable at time t. Positive coefficients indicate that a high value of a property leads to an increase of another property, while negative coefficients indicate that a high value of a property leads to an decrease of another property. To reveal positive and negative influence relations between properties within and across different networks, we visualize them as graphical influence network.

Table 1: Social and content network properties

Network type	Property	Description
Social	#Followers	The number of followers a user v has on a specific time point t.
Social	#Followees	The number of followees a user v has on a specific time point t.
Content	#Tweets	The number of tweets a user v has authored on a specific time point t.
Content	Hashtag ratio	The number of hashtags used by a user v on a specific time point t, normalized by the number of daily tweets authored by him/her.
Content	Retweet ratio	The number of messages authored by a user v which were retweeted by others on a specific time point t, normalized by the number of tweets he/she published that day.
Content	Retweeted ratio	The number of retweets produced by a user v on a specific time point t, normalized by the number of tweets user v published that day.

5. RESULTS

Our results reveal interesting influence patterns between social networks and content networks. The influence network in figure 1 shows the correlations detected in the multilevel regression analysis via arrows that point out influences between a property at time t and another property at time $t+1$.

The influence network reveals significant influences of social properties on content network properties. The strongest positive effects can be observed between the number of followers of a user and the content network properties - i.e., users' number of followers positively influences their link ratio, their retweeted ratio and their number of tweets. This indicates that users start providing more tweets and also more links in their tweets if their number of followers increases. Not surprisingly, users' tweets are also more likely to get retweeted if their number of followers increases, because more users are potentially reading their tweets.

Further, Figure 1 shows that the number of followees of the social network has positive and negative influences on the content network in our random dataset. While the positive effects point to the link and hashtag ratio, the negative effects point to the number of tweets and the retweeted ratio. This suggests that users who start following other users also

start using more hashtags and links. One possible explanation for this is that users get influenced by the links and hashtags used by the users they follow and might therefore use them more often in their own tweets. The negative effect of the number of followees on the number of tweets and the retweeted ratio suggests that users who start following many other users start behaving more like passive readers rather than active content providers.

Another observation of our experiment is that all properties influence themselves positively, which indicates that users who are active one day, tend to be even more active the next day. This indicates for example, that users who attract new followers one day tend to attract more new followers the day after.

6. CONCLUSIONS AND FUTURE WORK

The main contributions of this paper are the following: (i) We applied multilevel time series regression models to one selected Twitter datasets consisting of social and content network data and (ii) we explored influence patterns between social and content networks on Twitter. In our experiments we studied how the properties of social and content networks co-evolve over time. We showed that the adopted approach allows answering interesting questions about how users' behavior on Twitter evolves over time and the factors that impact this evolution. While our results are limited to the dataset used, our work illuminates a path towards studying complex dynamics of network evolution on systems such as Twitter. Our analyses may also facilitate social media hosts to promote certain features of the platform and steer users and their behavior. For example, one can see that from our analysis that usage of content features, such as hashtags and links, is highly influenced by social network properties such as the number of followers of a user. Therefore, social media hosts could try to encourage users to use more content features by introducing new measures such as a friend recommender techniques which might impact the social network of users. However, further work is warranted to study these ideas.

Overall, our findings on one small Twitter datasets suggest that there are manifold sources of influence between social and content network properties. Our results indicate that users' behavior and the co-evolution of content and social networks on Twitter is driven by social factors rather than content factors. Previous research by Anagnostopoulos et al. [1] showed that content on Flickr is not strongly influenced by social factors. This may suggest that different social media applications may be driven by different factors. The experimental setup used in our work can be applied to different datasets to study these questions in the future. Nevertheless, further work is required to confirm or refute this observations on other, larger datasets.

Our experiments also suggest that the number of followers strongly influences properties of the content network, which we interpret to mean that the number of followers is a very important motivation for Twitter users to add more content and use more content features like hashtags, URLs or retweets.

/* not sure if previous sentence is correct english (wac) */

However, the number of users a user is following can also have a negative influence on content network properties as one can see from figure 1. Our results suggests that an increase of a user's followees (i.e., the number of users he/she follows) implies that the user starts tweeting less and that his/her tweets get less frequently retweeted.

/* negative influence erwaehnst du heir das erste mal oder? in der results section steht das net... (wac) */

Further, our findings show that all properties influence themselves positively. This does not mean that the values of all properties always increase over time, but that they tend to increase depending on how much they increased the day before. For example, a Twitter user who started posting more links at day t, is likely to post even more links at day $t+1$ or a user who gain new followers at day t is likely to gain even more new followers at day $t+1$.

/* der absatz davor klingt nach copy and paste (wac) */

To summarize, our work highlights the existence of interesting influence relationships between content and social network on Twitter, and shows that multilevel time series regression analysis can be used to reveal such relationships and to study how they evolve over time. Based on the techniques developed by Wang and Groth [9], our work investigated influence patterns in a new domain, i.e. on microblogging platforms like Twitter. Our results are relevant for researchers interested in social network analysis, text mining and behavioral user studies, as well as for community hosts who need to understand the factors that influence the evolution of their users in terms of their content-related and social behavior.

Acknowledgments

This work is in part funded by the FWF Austrian Science Fund Grant I677 and the Know-Center Graz. Claudia Wagner is a recipient of a DOC-fForte fellowship of the Austrian Academy of Science.

7. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In Y. Li, B. Liu, and S. Sarawagi, editors, *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*, pages 7–15. ACM, 2008.

[2] A. Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873, July 2008.

[3] J. M. Gottman. *Time-Series Analysis: A Comprehensive Introduction for Social Scientists*. Cambridge University Press, 2009.

[4] A. F. Hayes. A primer on multilevel modeling. *Human Communication Research*, 32(4):385–410, 2006.

[5] G. Kitagawa. *Introduction to Time Series Modeling (Chapman & Hall/CRC Monographs on Statistics &*

Applied Probability). Chapman and Hall/CRC, 2010.

[6] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples (Springer Texts in Statistics)*. Springer, 2006.

[7] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall/CRC, 2004.

[8] C. Wagner and M. Strohmaier. The wisdom in tweetonomies: Acquiring latent conceptual structures from social awareness streams. In *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, april 2010.

[9] S. Wang and P. Groth. Measuring the dynamic bi-directional influence between content and social networks. In P. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Pan, I. Horrocks, and B. Glimm, editors, *The Semantic Web Ű ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 814–829. Springer Berlin / Heidelberg, 2010.