# Content-Based Discovery of Twitter Influencers

Chiara Francalanci, Irma Metra

Department of Electronics, Information and Bioengineering
Polytechnic of Milan, Italy
irma.metra@mail.polimi.it
chiara.francalanci@polimi.it

## Abstract

Identifying social media influencers in a given domain is considered key to building a brand's reputation. Influencers are opinion makers who play a critical role in determining the dynamics with which information spreads across a social network. In Twitter, a large number of followers is considered a fundamental indicator to discover influencers. The assumption is that a user with a large number of followers has a large audience and, thus, is more likely to influence the opinion of people in any given domain. Our claim is that influencers can exert an influence only when the content that they share is considered interesting by their followers. In this paper, we propose a content-based measure of influence, called COAX that includes, but is not limited to the number of followers. COAX is tested on a sample of over 10.000 users from random domains according to the Analytic Hierarchy Process (AHP). Preliminary results show how COAX can provide a ranking that is significantly different from that obtained by means of the number of followers alone.

**Keywords:** social media; influencers; influence; Twitter.

## 1. Introduction

Identifying social media influencers in a given domain is considered key to building a brand's reputation (Bruni L. , 2014). Influencers are opinion makers who play a critical role in determining the dynamics with which information spreads across a social network. In Twitter, a large number of followers is considered a fundamental indicator to discover influencers. The assumption is that a user with a large number of followers has a large audience and, thus, is more likely to influence the opinion of people in any given domain.

Our claim is that influencers can exert an influence only when the content that they share is considered interesting by their followers. As a consequence, they are influential in selected domains where they have the capability to share interesting content. The previous academic literature supports our claim by showing how a variety of variables describing content can have an impact on the probability with which content itself is shared. For example, in (Bruni, Francalanci, & Giacomazzi, 2013) the authors claim that linking multimedia content in a Tweet increases the average number of retweets. In (Boyd, Golde, & Lotan, 2010) authors note that a content that has had an impact on a user's mind is shared. In (Suh, Hong, Pirolli, & Chi, 2010) authors discover how most content is retweeted only once and (Ota, Maruyama, & Terada, 2012) introduces the concept of *depth* of retweets to measure the impact of the original tweet. Overall, the academic literature is heading towards the concept of *influence*, i.e. the actual impact that a tweeter has on his audience and on other users that they are not directly connected with.

In this paper, we propose a content-based measure of influence, called COAX that includes, but is not limited to the number of followers. The number of followees, favorites, tweets, listed, mentions, urls, hashtags, retweets and favorited are considered in conjunction with the more traditional number of followers. COAX is tested on a sample of over 10.000 users from random domains according to the Analytic Hierarchy Process (AHP). Preliminary results show how COAX can provide a ranking that is significantly different from that obtained by means of the number of followers alone. They also show that the methodology is very robust, as proved by the sensitivity analysis.

## 2. COAX: A framework to build Influence Metrics with AHP

### Introduction

Our starting point has been the collection of a dataset of 11466 Twitter active users. The information gathered for each of them is the **number** of followers, following (the people they are following), lists he is member/owner of, tweets and tweets the user marked as favorite; for each user we further collected the **number** of retweets of his last 100 tweets, times tweets were marked as favorite, hashtags used, url's used and people the user mentioned.

Our goal was to provide a ranking algorithm based on this information, which would help identifying influencers or important tweeters. (Metra, 2014) In order to achieve this goal we used the Analytical Hierarchy Process (Saaty, 1980) to provide us with specific weights for each parameter of the data collected. AHP requires a specific problem setup composed of two steps: 1) define **the aggregation criteria**, i.e. divide variables in categories and subcategories as shown in Fig.1; 2) define **the objective**, i.e. discover appropriate weights for our parameters related to Tweeter activity.

### *1. Variable operationalization and aggregation criteria*

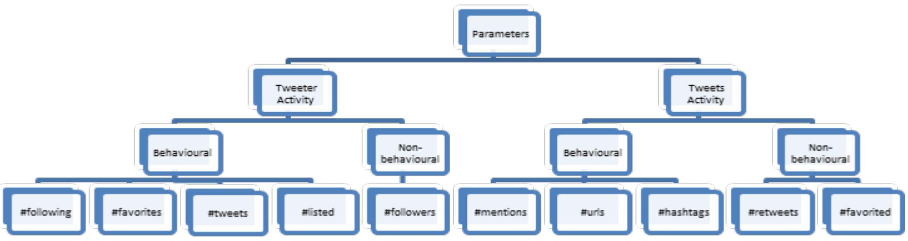


**Fig. 1**. Hierarchical aggregation of variables.

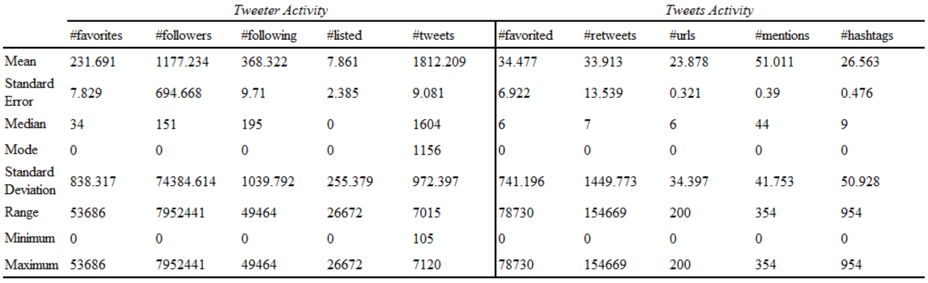
We decided to use 10 parameters based on tweeter and tweet activity. Our attention is limited to active tweeters, meaning people who use twitter actively, almost on a daily basis. Studies have shown that having a frequency of two or less tweets per day makes a user be not a spammer but an active one and most probably a person who is quite influential. In our dataset, the last time the user was active in twitter by posting a tweet is at least 120 days. We have selected the following categories:

**Tweeter Activity**: This category includes all the parameters that are totals (sum) related to the tweeter such as number of favorite tweets (#favorites), number of followers (#followers), number of people that the tweeter is following (#following), number of lists the tweeter is owner of/member of (#lists) and total number of tweets posted from the tweeter (#tweets). **Tweets Activity**: This category includes all the parameters which were obtained as a sum for the last 100 tweets of each tweeter. It is composed of the total number of how many times the last 100 tweets of the tweeter have been marked as favorite by other tweeters (#favorited), total number of how many times the last 100 tweets of the tweeter have been retweeted (#retweets), total number of URL's used in the last 100 tweets (#urls), total number of people mentioned in the last 100 tweets (#mentions) and total number of hashtags used in the last 100 tweets (#hashtags).These two categories are also divided both in two subcategories:

**Behavioral**: Behavioral parameters are considered the ones that are a consequence of the tweeter's actions. **Non-behavioral**: Non-behavioral parameters are considered those which are a consequence other tweeters' actions with respect to a specific tweeter.

### 2. *Data sample*

The final dataset to test COAX contains 11.466 active users; we used 9000x2 API calls, 2000 API calls less then what we had calculated in the worst case. Data collection started 22/01/2014 at 18:15 and ended 24/10/2014 at 17.30; so a total of approximately 47 hours, 7 hours less than the worst case. The following tables summarize descriptive statistics regarding the Tweeter and Tweet activity of the 11466 users.



| | *Tweeter Activity* | | | | | *Tweets Activity* | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | #favorites | #followers | #following | #listed | #tweets | #favorited | #retweets | #urls | #mentions | #hashtags |
| Mean | 231.691 | 1177.234 | 368.322 | 7.861 | 1812.209 | 34.477 | 33.913 | 23.878 | 51.011 | 26.563 |
| Standard Error | 7.829 | 694.668 | 9.71 | 2.385 | 9.081 | 6.922 | 13.539 | 0.321 | 0.39 | 0.476 |
| Median | 34 | 151 | 195 | 0 | 1604 | 6 | 7 | 6 | 44 | 9 |
| Mode | 0 | 0 | 0 | 0 | 1156 | 0 | 0 | 0 | 0 | 0 |
| Standard Deviation | 838.317 | 74384.614 | 1039.792 | 255.379 | 972.397 | 741.196 | 1449.773 | 34.397 | 41.753 | 50.928 |
| Range | 53686 | 7952441 | 49464 | 26672 | 7015 | 78730 | 154669 | 200 | 354 | 954 |
| Minimum | 0 | 0 | 0 | 0 | 105 | 0 | 0 | 0 | 0 | 0 |
| Maximum | 53686 | 7952441 | 49464 | 26672 | 7120 | 78730 | 154669 | 200 | 354 | 954 |

**Table 1.** Tweeter and Tweets Activity Descriptive Statistics

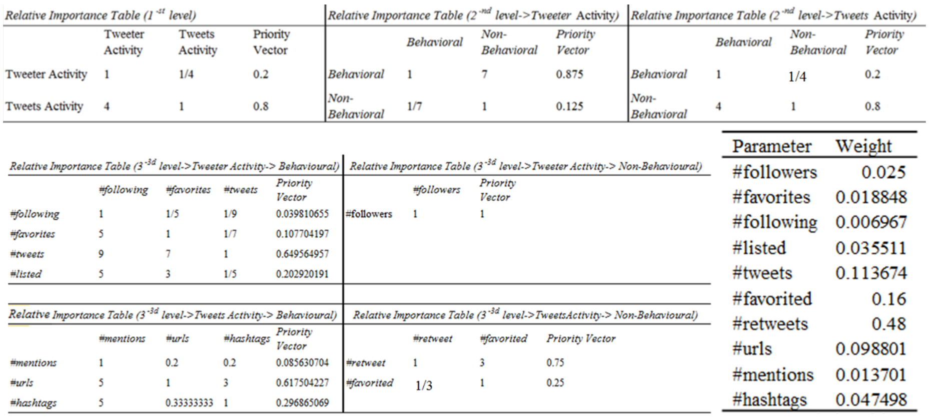
## Compare Criteria

### *Relative Importance Table and Priority Vector Calculation*

In order to build relative importance tables must be calculated, we needed to compare our criteria in a pairwise fashion. Pairwise comparisons are quantified by using a scale, which is a one-to-one mapping between the set of discrete linguistic choices available to the decision maker and a discrete set of numbers which represent the importance, or weight, of the previous linguistic choices. According to this scale, the available values for the pairwise comparisons are members of the set: {9, 8, 7, 6, 5, 4, 3, 2, 1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, 1/8, 1/9} (Saaty, 1980)

The next step is to extract the relative importance's implied by the comparisons (see Table 2.) Saaty asserts that to address this problem one has to estimate the right principal eigenvector of the previous matrices. Given a judgment matrix with pairwise comparisons, the corresponding maximum left eigenvector is approximated by using the geometric mean of each row.

*Relative Importance Tables and final priority vector of Tweeter and Tweets Activity Parameters*



*Relative Importance Table ($1^{-st}$ level)*

| | Tweeter Activity | Tweets Activity | Priority Vector |
|---|---|---|---|
| Tweeter Activity | 1 | 1/4 | 0.2 |
| Tweets Activity | 4 | 1 | 0.8 |

*Relative Importance Table ($2^{-nd}$ level->Tweeter Activity)*

| | *Behavioral* | *Non-Behavioral* | *Priority Vector* |
|---|---|---|---|
| *Behavioral* | 1 | 7 | 0.875 |
| *Non-Behavioral* | 1/7 | 1 | 0.125 |

*Relative Importance Table ($2^{-nd}$ level->Tweets Activity)*

| | *Behavioral* | *Non-Behavioral* | *Priority Vector* |
|---|---|---|---|
| *Behavioral* | 1 | 1/4 | 0.2 |
| *Non-Behavioral* | 4 | 1 | 0.8 |

*Relative Importance Table ($3^{-3d}$ level->Tweeter Activity-> Behavioural)*

| | *#following* | *#favorites* | *#tweets* | *Priority Vector* |
|---|---|---|---|---|
| *#following* | 1 | 1/5 | 1/9 | 0.039810655 |
| *#favorites* | 5 | 1 | 1/7 | 0.107704197 |
| *#tweets* | 9 | 7 | 1 | 0.649564957 |
| *#listed* | 5 | 3 | 1/5 | 0.202920191 |

*Relative Importance Table ($3^{-3d}$ level->Tweeter Activity-> Non-Behavioural)*

| | *#followers* | *Priority Vector* |
|---|---|---|
| #followers | 1 | 1 |

*Relative Importance Table ($3^{-3d}$ level->Tweets Activity-> Behavioural)*

| | *#mentions* | *#urls* | *#hashtags* | *Priority Vector* |
|---|---|---|---|---|
| *#mentions* | 1 | 0.2 | 0.2 | 0.085630704 |
| *#urls* | 5 | 1 | 3 | 0.617504227 |
| *#hashtags* | 5 | 0.33333333 | 1 | 0.296865069 |

*Relative Importance Table ($3^{-3d}$ level->TweetsActivity-> Non-Behavioural)*

| | *#retweet* | *#favorited* | *Priority Vector* |
|---|---|---|---|
| *#retweet* | 1 | 3 | 0.75 |
| *#favorited* | 1/3 | 1 | 0.25 |

| Parameter | Weight |
|---|---|
| #followers | 0.025 |
| #favorites | 0.018848 |
| #following | 0.006967 |
| #listed | 0.035511 |
| #tweets | 0.113674 |
| #favorited | 0.16 |
| #retweets | 0.48 |
| #urls | 0.098801 |
| #mentions | 0.013701 |
| #hashtags | 0.047498 |

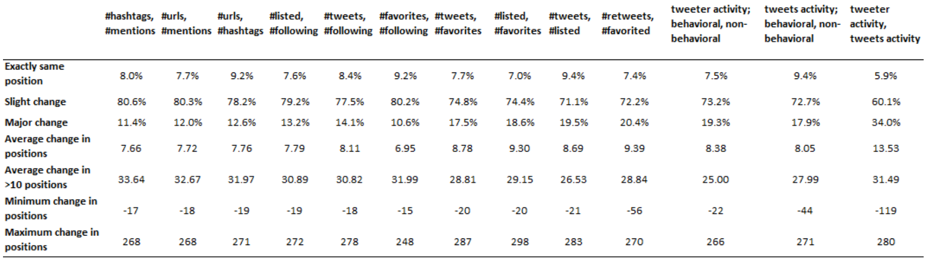
**Table 2.** Relative Importance Tables, the first second and third level

These tables represent the relative importance tables in the first, second and third level. The values of importance have been decided by the authors of this paper. A detailed discussion can be found in (Metra, 2014). To get the final ranking, for each tweeter we perform the weighted sum of the parameters, based on the Weight Vector presented in the previous diagram.

### Results and Sensitivity Analysis

We performed a sensitivity analysis of our judgments integer values by decreasing them by one (so we have more than 10% tweak in the parameter value), one at a time for each judgment done with respect to our 10 parameters. That means that at the third level we performed ten changes.

Tweaking parameters maintains the stability of the rankings (see Table 3). In all the cases, the percentage of no changes and slight changes reaches a minimum of 66% by considering also the first level of changes, which as earlier described introduces a larger amount of changes. Throughout all the other levels, this percentage is always greater than 80%.



| | #hashtags, #mentions | #urls, #mentions | #urls, #hashtags | #listed, #following | #tweets, #following | #favorites, #following | #tweets, #favorites | #listed, #favorites | #tweets, #listed | #retweets, #favorited | tweeter activity; behavioral, non-behavioral | tweets activity; behavioral, non-behavioral | tweeter activity, tweets activity |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Exactly same position | 8.0% | 7.7% | 9.2% | 7.6% | 8.4% | 9.2% | 7.7% | 7.0% | 9.4% | 7.4% | 7.5% | 9.4% | 5.9% |
| Slight change | 80.6% | 80.3% | 78.2% | 79.2% | 77.5% | 80.2% | 74.8% | 74.4% | 71.1% | 72.2% | 73.2% | 72.7% | 60.1% |
| Major change | 11.4% | 12.0% | 12.6% | 13.2% | 14.1% | 10.6% | 17.5% | 18.6% | 19.5% | 20.4% | 19.3% | 17.9% | 34.0% |
| Average change in positions | 7.66 | 7.72 | 7.76 | 7.79 | 8.11 | 6.95 | 8.78 | 9.30 | 8.69 | 9.39 | 8.38 | 8.05 | 13.53 |
| Average change in >10 positions | 33.64 | 32.67 | 31.97 | 30.89 | 30.82 | 31.99 | 28.81 | 29.15 | 26.53 | 28.84 | 25.00 | 27.99 | 31.49 |
| Minimum change in positions | -17 | -18 | -19 | -19 | -18 | -15 | -20 | -20 | -21 | -56 | -22 | -44 | -119 |
| Maximum change in positions | 268 | 268 | 271 | 272 | 278 | 248 | 287 | 298 | 283 | 270 | 266 | 271 | 280 |

**Table 3.** Sensitivity Analysis, Summary Table

## 3. Discussion and Conclusions

COAX introduces compelling tasks and research, as how to address in a proper way an influence metric. In particular the results enforce the need for an influence metric that is prescriptive, comprehensive, general and mathematical.

These results have an impact on the academic literature since they provide an innovative methodology to calculate and parameterize influence. Such results fill in the gap that existed until now in the research with respect to influence and introduces further research challenges. They also have an impact on practitioners since it provides a disclosed framework, applicable to any domain, very easy to implement, robust and more reliable.

A fundamental result of COAX was the fact that the ranking proposed is much different from the ranking based on the number of followers. Major changes in ranking happen when the judgments in the first level change, as the weights of AHP are more sensitive to them. We need to have very precise judgments especially regarding the first level of the hierarchy of our parameters.

## 4. References

Boyd, D., Golde, S., & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *IEEE*, 1-10.

Bruni, L. (2014, March). A Methodological Framework to Understand and Leverage the Impact of Content on Social Media Influence. Italy: Politechnic of Milan.

Bruni, L., Francalanci, C., & Giacomazzi, P. (2013). Measuring the Web reputation impact of events: preliminary evidence from a city brand listening project. *ENTER*. Innsbruck: Springer International Publishing.

Metra, I. (2014, April). Influence based exploration of Twitter Social Network. Italy. Retrieved from http://hdl.handle.net/10589/92709

Ota, Y., Maruyama, K., & Terada, M. (2012). Discovery of Interesting Users in Twitter by Overlapping Propagation Paths of Retweets. *International Conferences on Web Intelligence and Intelligent Agent Technology, 2012 IEEE/WIC/ACM International Conferences. III*, pp. 274 - 279. Macau: IEEE.

Saaty, T. L. (1980). *The Analytic Hierachy Process*. New York, NY, USA: McGraw-Hill International.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. *Social Computing (SocialCom), 2010 IEEE Second International Conference* (pp. 177 - 184). Minneapolis, MN: IEEE.