Testing Social Network Metrics for Measuring Electoral Success in the Italian Municipal Campaign of 2011

Paolo Ceravolo
Computer Science Department
Università degli Studi di Milan,
Via Bramante 65, 26013 Crema (CR), Italy
Email: paolo.ceravolo@unimi.it

Stefano Guerretti
Computer Science Department
Università degli Studi di Milan,
Via Bramante 65, 26013 Crema (CR), Italy
Email: stefano.guerretti@studenti.unimi.it

***Abstract*—It is often argued that the bias hidden in Social Media data prevent from using them for any statistical inference. In this paper, we investigate the practicability of a new method for predicting electoral outcomes that is less affected by demographics and self-selection bias. In particular, we put in place a first test to understand which social network analysis metrics can exhibit positive correlation with electoral success. Our analysis is not intended to use social media audience as a sample of the whole electorate but just as a sample of the supporters of a candidate. In conclusion, we speculate on the information we can extract measuring the social network of the groups of supporters. Essentially, we get an overview on the variety and extent of the segments of the population represented in these groups, and this probably correlates with the capacity to attract consensus.**

***Keywords*-Social Media Bias; Social Network Analysis; Predicting Electoral Outcomes; Italian Municipal Campaign of 2011.**

I. Introduction

Everyday social media provide a formidable trail of human activities and preferences. The magnitude of records and the detail level of the information produced has instigated the fascinating idea that it is possible to predict real world outcomes from the analysis of social media. One of the most attractive area for testing the issue is related to electoral predictions, where we have both high public attention and empirical proof of the validity of the predictions.
But the feasibility of this idea is controversial. Despite the fact that in the last years several studies focused on this issue and most of them reported positive results [1], [19], [9], [8], [16], [14], [2], we lack of consolidated approaches to be replicated in order to gain statistical evidence of the validity of these predictions. In particular, several authors have criticized current studies because of a superficial or weak application of the scientific method [13], [11], [6]. The most relevant deficiencies reported are:

- Lack of a standard method; for example in measuring user preferences and opinions or in testing predictions on a valid baseline.
- Demographics bias is disregarded. Even if it is well known that social media are not a random sample of the population; for instance not every active user is eligible to vote, and geographical fitting cannot be controlled.
- Self-selection bias is ignored. Even if it is well known that there is a silent majority that is inactive, or some controversial issues can get more attention than others that also impact on voter's deliberation.

Our aim with this paper is to investigate the practicability of a new method for predicting electoral outcomes that is not affected by demographics and self-selection bias. Our idea is that the analysis of the social network of the supporters of a candidate can give information on that candidate's ability to attract consensus from the electorate. Hence, our analysis is not intended to use social media audience as a sample of the whole electorate but just as a sample of the supporters of a candidate. In particular, our idea is to consider Social Network Analysis (SNA) metrics to evaluate the variety of population segments represented in the group of supporters. Therefore, the hypothesis we have to verify is if we are allowed to assume the correlation between some properties of the social network of the supporters groups and the electoral outcome.
Our study was conducted on a selection of local elections carried out during the italian municipal campaign of 2011. We identified all the cases where Facebook groups of supporters of at least the two most important candidates were available and we analyzed the social network of these groups. Our analysis was performed first by constructing the adjacency matrix of the groups of supporters. Then we calculated seven common social network metrics for verifying if the orientation of some of these metrics constantly recurs in association with the group of supporters of the elected candidate. In our study the *Connected Components* and the *Average Path Length* metrics presented a constant positive orientation in the groups supporting the elected candidates, while the *Graph Density* presented a constant negative orientation in the groups supporting the elected candidates.

We conclude that the approach tested is promising and it is eligible to be applied in a deeper investigation. Due to the limited number of cases analyzed[1], we cannot claim that our study provides any specific evidence on the validity of the method proposed. Nevertheless we consider our research is an original contribution, especially when it identify a new approach that can encompass the limitations of the current research approaches. In particular, by moving the focus from using social media as a sample of the whole population, using them as a sample of the candidate's supporters. This way the group of supporter can be studied from the point of view of its components.

The paper is structured in the following way. In section II we provide an overview of the state of the art on the topic. In section III we describe the research method we applied, the data collected and the analysis performed. In section IV we go to conclusions.

II. Related Work

The connections between social media and political orientation was investigated with increasing intensity in the last years. In [1] the authors have found that linkage patterns among bloggers reflect the blogosphere along party lines, and this suggested that the users' activities on social media can be considered as a reflection of the political orientation of a population. In [17] the authors proposed an analysis on the 2007 French presidential election, concluding that the simple count of candidates' mentions in the press was a better predictor of electoral success than many election polls. Williams and Gulati deeply investigated the correlation among the number of supporters generated on Facebook and other indicators relevant to the U.S. elections [18], [19], [20]. For instance, they observed that, in general, Democrats have more supporters than Republicans, incumbents more than challengers, an higher number of college educated citizen in the electoral district is positively correlated to the number of candidates' supporters, as well as the quality of the activity produced on the Facebook page. In [8] and [9] social network analysis is applied to rank candidates in a political election. The method proposed uses Google to construct a network of the mentions related to candidates and assign a score based on the betweenness of the interconnected website containing these mentions. In [16] the authors analyzed the German federal election in 2009. Based on the results obtained, they claim that the mere number of tweets reflects voter preferences with accuracy close to the election polls made during this campaign. However, there is also literature which questioned the validity of the prediction models proposed. In [6], Gayo-Avello proposes a survey on studies about the predictive power of twitter data underlining several limitations in the research methods adopted in the revised literature, in particular due to lack of estimators for the bias between the true electorate and the data mined from social media. However, certain studies propose methodologies to set up estimators. In [2] and [15] the authors underline that merely counting the tweets is not sufficient for electoral predictions and propose to improve the quality of data collection by performing sentiment analysis. This way it is possible to distinguish generic interest from explicit consensus. In [4] the focus is on exploiting socio-demographics and census information to correct the bias in the online data. In [12] the authors propose a general method for assessing the relevance of the expressed opinions according to the different behaviors adopted by users in generating content, which expresses different level of engagement.

[1]And to the fact that we limited our investigation to a single campaign were the political orientation of the electorate was clearly determinate.

III. Research Methodology

As explained, the current research brings attention to the need for handling the bias of social media data. In order to contribute to exploring approaches encompassing this problem, our proposal is to consider the social media users not as a sample of the population but as a sample of the supporters of a candidate. Therefore, we are not interested in building predictions based on the mere dimension of this group, whereas the aim is mining the structure of the group to have an estimation on the spread over the whole population. Thus, we propose to focus on the social network of the supporters of a candidate to obtain information on his/her capacity to attract consensus from the whole electorate.

A. Research Challenge

In particular, our aim is to put in place a first test to understand which social network analysis metrics can exhibit a positive correlation with the electoral success. The procedure we followed includes: the identification of a group of supporters, the construction of the social network of this group, the evaluation of a set of metrics, the correlation analysis among the orientation of these metrics and the electoral success. The set of metrics we selected for our test included: Average Degree (AD), Network Diameter (ND), Graph Density (GD), Modularity (M), Connected Components (CC), Avg. Clustering Coefficient (ACC), Avg. Path Length (APL) [3].

B. Data Collection

To identify the group of supporters of a candidate we are required to distinguish among the users that are connected to the candidate through the social media simply because of the attention generated by the candidate[2] or because of a specific interest in supporting the candidate. As seen in section II, sentiment analysis techniques can be adopted for mining the orientation of users. But because the accuracy achieved by these techniques is debatable, we decided to adopt a more straightforward approach. Our analysis focused on Facebook

[2]Note that attention can be both positively or negatively oriented.

groups that by definition collect users willing to cooperate on a common task[3]. When accepted as a member of a group it is possible to use the Facebook graph API to access the list of the members of a group and then their friends with a degree that depend on the policy settings of each user [4]. This way we can construct the adjacency matrix representing the social network of a group[5] and calculate the metrics listed in section III-A based on these data.

Our survey was conducted during the Italian Municipal Campaign of 2011. Focusing on a Municipal Campaign we had the opportunity to test several cases at the same time. Moreover, as the approach we followed focused on Facebook groups of supporters, we were required to restrict our tests to cases where Facebook groups of supporters of the two most important candidates were available. We then identified two primary elections: Crema and Palermo, and three elections for office: Milan, Naples and Turin [6].

C. Data Analysis

The dimension of the data under analysis does not allow us to infer any information on the coherence among the results of our metrics and the political consensus. For this reason we limit our analysis testing different metrics as predictors of the final outcome.

For each election we collected the results of the tested metrics as reported in table I, that shows the case of Milan.

Table I
MILAN, MUNICIPAL ELECTION 2011

Candidate	SNA Metrics						
	AD	ND	GD	M	CC	ACC	APL
Pisapia	16,76	8	0,019	0,345	98	0,353	2,798
Moratti	61,21	5	0,134	0,157	35	0,496	2,014

To compare these data we have to apply a standardization procedure. All these different metrics span over different numerical ranges, moreover the scale (size and density) of the networks analyzed impact on these ranges. For this reason we implemented a procedure involving a non-linear transformation and a statistical normalization, following the approach proposed in [10]. The normalized values of the tested metrics are shown in table II.

Table III reports the results we have obtained in the five tested cases. When the greater value of a SN metric is associated to the elected candidate we report 1, if not we report 0. As It can be observed the *Connected Components* and the *Average Path Length* metrics present a constant positive orientation in the groups supporting the elected candidates, while the *Graph Density* presents a constant negative orientation in the groups supporting the elected candidates. In our opinion these results can be explained as indicators of the variety in the community of supporters that in turn may be proportional to the width of impact of a candidate on the whole electorate. With *Connected Components* we have an indicator of how many components in the network have more connections inside the component than outside. The *Average Path Length* is an indicator of the distance among the nodes belonging to a network and has grater values with networks divided in weakly connected areas. The *Graph Density* is a measure of the ratio of the number of connections and the number of possible connections, so we have lower values when the network is divided in different sub-communities. If we consider that in social network analysis it is common to interpret the similarity between nodes based on the number of shared neighbors[7], we understand that all these metrics express the variety within the community of supporters. A visual representation of the social network of supporters for the two candidates to the office in Milan is depicted in fig. 1 and 2. The network is displayed graphically as a set of nodes with different colors and sizes. The node's size depends on the Betweenness Centrality score [5], i.e. the number of nodes directly or indirectly connected to individual nodes. In detail, the networks depicted show two different situations: in the Pisapia's case we have a single node that joins all the others, emphasizing its ability to communicate with the entire group, as a leader; in the Moratti's case we have two nodes above the average that are the two pillars along which communications move between individual nodes. The color is decided by the Connected Component score, each component is identified by a different color that represents a different community in support of the candidate. A simple visual analysis of these networks shows that we have a quite clear division of components supporting the candidate Moratti, because the colors are limited and nodes grouped around colors, while the picture for the candidate Pisapia is more confusing, due to the large number of components (it

Table II
MILAN, MUNICIPAL ELECTION 2011

Candidate	SNA Metrics						
	AD	ND	GD	M	CC	ACC	APL
Pisapia	0,18	0,08	0,01	0,04	0,80	0,01	0,03
Moratti	0,23	0,10	0,02	0,03	0,70	0,02	0,04

[3]Facebook groups are similar to Facebook pages but contain a different set of features. Groups are a way of enabling a number of people to come together online to share information and discuss specific subjects. The administrators can take a group close or open. Depending on this the membership is up for approval or not and the content posted on the wall is restricted to members or not.

[4]Facebook's privacy policy changed many times over the years, for this reason there is no assurance the process we followed will be reproducible in the future.

[5]In particular this was obtained by running the Netvizz application

[6]More details on these elections are available at http://en.wikipedia.org/wiki/Italian_local_elections,_2011

[7]Where the notion of neighbor can be parametrized with the relations to be followed (usually friendship), the number of relations to be considered (k-nearest-neighbor) and the number of hops accepted (typically two options are considered: fiends or fiends-of-friends).

is often hard to perceive the difference between one color and the other) and the presence of links between them.

Table III
TESTS ON SOCIAL NETWORK ANALYSIS METRICS.

Tested Election	SNA Metrics						
	AD	ND	GD	M	CC	ACC	APL
Crema	1	1	0	1	1	0	1
Palermo	1	1	0	0	1	1	1
Milan	0	1	0	1	1	0	1
Naples	0	0	0	0	1	0	1
Turin	0	1	0	1	1	0	1

IV. CONCLUSIONS

Due to the approach followed and to the source of data selected (Facebook groups) the number of cases analyzed does not allow us to consider this test statistically significant [8]. In addition the data collected are affected by the specific settings in the Facebook privacy policies of group's members. We are perfectly aware of these limitations but we considered this test as an useful clue that indicate one promising track to be followed in future studies.

We also think that this work should be considered as a contribution to the broad variety of studies that are proposing novel approaches to deal with the bias hidden in social media data. In particular we conjecture that relevant information could be obtained by analyzing the social network of a group of users expressing preferences and opinions. In effect, the social network can tell us if one group is self-referential or not, if it is representative of a restricted number of segments of the population or if it represents a variety of different segments.

This research line needs deeper investigation in at least two directions. On one side it is necessary to test this method on a wide set of examples, to get statistical evidence. On the other side it necessary to improve the quality of the analysis made on the social network, following, for instance, approaches based on consensus emergence [7].

REFERENCES

[1] Adamic, L. A. and Glance, N. *The political blogosphere and the 2004 US election: Divided they blog*. In Proceedings of the 3rd International Workshop on Link Discovery, 36-43, 2005.

[2] Bermingham, A. and Smeaton, A.F. *On using Twitter to monitor political sentiment and predict election results*. In Proceedings of the Sentiment Analysis where AI meets Psychology Workshop at IJCNLP, 2011.

[3] Brinkmeier, M. and Schank, T. *Network Statistics*. Network Analysis Lecture Notes in Computer Science, vol. 3418, pp. 293-317, 2005.

[8] For instance a Cochran's Q test gives a result of 4,39, that is far short of the threshold.

[4] Choy, Murphy et al. *US Presidential Election 2012 Prediction using Census Corrected Twitter Model.* arXiv preprint arXiv:1211.0938, 2012.

[5] Freeman, Linton. *A set of measures of centrality based on betweenness*. Sociometry, vol. 40, pp. 3541, 1977.

[6] Gayo-Avello, D. *I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper*. CoRR, vol. 1204.6441, http://arxiv.org/abs/1204.6441. 2012.

[7] Gianini, G. Damiani, E. Ceravolo, P. Consensus emergence from naming games in representative agent semantic overlay networks. In: On the move to meaningful internet systems: OTM 2008 workshops: OTM confederated international workshops and posters, pp. 1066-1075, ISBN 9783540888741, 2008.

[8] Gloor, P. A. et al. *Web science 2.0: Identifying trends through semantic social network analysis.* Computational Science and Engineering, 2009. CSE'09. International Conference on. Vol. 4. IEEE, 2009.

[9] Grippa, F. and Del Vecchio, P. *Take me to your Leader: Predicting Political Leadership using Social Network Metrics*, in Proceedings of SUNBELT, 2008.

[10] Joseph, A. and Chen, G. Composite Centrality: A Natural Scale for Complex Networks. In Signal Image Technology and Internet Based Systems (SITIS), Eighth International Conference on, pp. 739-743, IEEE, 2012.

[11] Jungherr, A. Jurgens, P., and Schoen, H. *Why the pirate party won the german election of 2009 or the trouble with predictions: A response to Tumasjan, a., Sprenger, t. o., Sander, p. g., & Welpe, i. m. "predicting elections with twitter: What 140 characters reveal about political sentiment"*. Social Science Computer Review, 2011.

[12] Chen, Lu, Wenbo Wang, and Amit P. Sheth. *Are Twitter Users Equal in Predicting Elections? A Study of User Groups in Predicting 2012 US Republican Presidential Primaries*. K. Aberer et al. (Eds.): SocInfo 2012, LNCS 7710, pp. 379392, 2012.

[13] Metaxas, P.T. Mustafaraj, E. Gayo-Avello, D. , *How (Not) to Predict Elections,* Privacy, security, risk and trust (passat), IEEE Third International Conference on Social Computing (socialcom), pp.165-171, 9-11 Oct. 2011 doi: 10.1109/PASSAT/SocialCom.2011.98

[14] O'Connor, B. Balasubramanyan, R. Routledge, B.R. and Smith, N.A. *From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series*, 4th International AAAI Conference on Weblogs and Social Media, May 23-26, 2010.

[15] Sang, E.T.K., Bos, J. *Predicting the 2011 Dutch Senate Election Results with Twitter*. In: Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks, pp. 5360, 2012.

[16] Tumasjan, A. Sprenger, T. O. Sandner, P. G. and Welpe I. M. *Predicting elections with twitter: What 140 characters reveal about political sentiment*. In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, pp. 178-185, 2010.

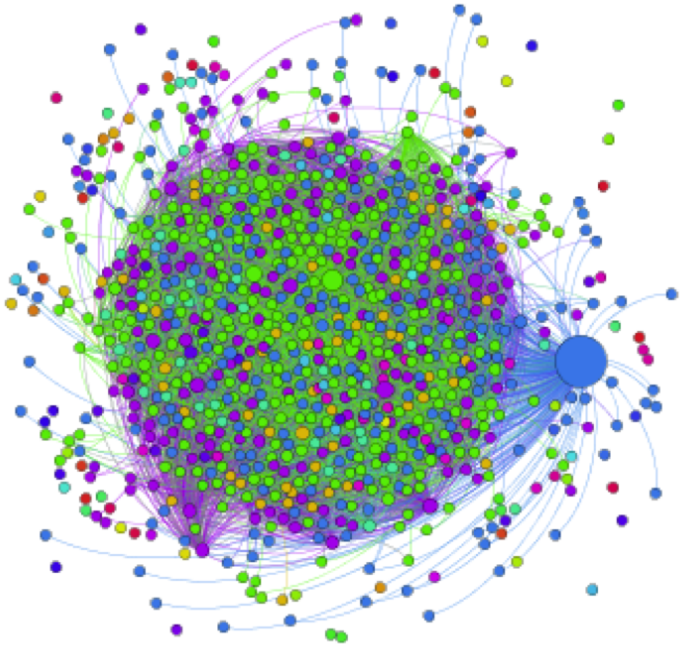

Figure 1. Visual Representation of the SN of the supporters of Pisapia.

Figure 2. Visual Representation of the SN of the supporters of Moratti.

[17] Véronis, J. *Citations dans la presse et résultats du premier tour de la présidentielle 2007*. Retrieved December 15, 2009 from http://aixtal.blogspot.com/2007/04/2007-la-presse-fait-mieux-que-les.html

[18] Williams, Christine B. and Gulati, Jeff, *Social Networks in Political Campaigns: Facebook and the 2006 Midterm Elections*, Annual Meeting of the American Political Science Association Chicago, Illinois, August 30 September 2, 2007.

[19] Williams, Christine B. and Gulati, Jeff, *What is a Social Network Worth? Facebook and Vote Share in the 2008 Presidential Primaries*. In Annual Meeting of the American Political Science Association, 1-17. Boston, 2008.

[20] Williams, Christine B. and Gulati, Jeff, *Social Networks in Political Campaigns: Facebook and Congressional Elections 2006, 2008*, APSA 2009 Toronto Meeting. SSRN: http://ssrn.com/abstract=1451451, 2009.