Multi-aspect Entity-centric Analysis of Big Social Media Archives

Pavlos Fafalios[1], Vasileios Iosifidis[1], Kostas Stefanidis[2], and Eirini Ntoutsi[1]

[1] L3S Research Center, University of Hannover, Germany
{fafalios, iosifidis, ntoutsi}@l3s.de
[2] Faculty of Natural Sciences, University of Tampere, Finland
kostas.stefanidis@uta.fi

Abstract. Social media archives serve as important historical information sources, and thus meaningful analysis and exploration methods are of immense value for historians, sociologists and other interested parties. In this paper, we propose an *entity-centric* approach to analyze social media archives and we define measures that allow studying how entities are reflected in social media in different time periods and under different aspects (like popularity, attitude, controversiality, and connectedness with other entities). A case study using a large Twitter archive of 4 years illustrates the insights that can be gained by such an entity-centric multi-aspect analysis.

1 Introduction

Social networking services have now emerged as central media to discuss and comment on breaking news and noteworthy events that are happening around the world. In Twitter, for example, every second around 6,000 tweets are posted, which corresponds to over 350,000 tweets per minute, 500 million tweets per day and around 200 billion tweets per year[3].

Such user-generated content can be seen as a comprehensive documentation of society and is therefore of immense historical value for future generations [7]. Although there are initiatives and works that aim to collect and preserve social media archives (e.g., the Twitter Archive at the Library of Congress [25]), the absence of meaningful access and analysis methods still remains a major hurdle in the way of turning such archives into useful sources of information for historians, journalists and other interested parties [7]. Besides, when exploring archived data, analysts are not interested in the documents per se, but instead they want to see, compare, and understand the behavior of (and trends about) entities, like companies, products, politicians, music bands, songs and movies, thus calling for entity-level analytics over the archived data [22].

In this paper, we propose an *entity-centric multi-aspect* approach to analyze social media archives. Our approach allows tracking of how entities are reflected in a collection of user-generated content (e.g., tweets) and how such information evolves over time and also with respect to other entities. Specifically, we

[3] http://www.internetlivestats.com/twitter-statistics/ (June 21, 2017)

define measures for the temporal analysis of an entity in terms of its: *popularity*, *attitude* (predominant sentiment), *sentimentality* (magnitude of sentiment), *controversiality*, and *connectedness* to other entities. A distinctive characteristic of our approach is that it does not rely on service-specific labels (like #hashtags and @mentions), but it exploits *entity linking* and thus can be applied over any type of time-annotated texts.

We examine the insights gained by the proposed measures by analyzing a large collection of billions of tweets spanning a period of 4 years. Such analytics enable to answer questions like:

- *How did the popularity of Greek Prime Minister, Alexis Tsipras, evolve in 2015? Were there any "outlier" periods, i.e., periods of extremely high or low popularity? What were the entities discussed in social media together with Alexis Tsipras during these periods?*
- *How did the predominant sentiment about Donald Trump and Hillary Clinton vary during 2016? Were there any controversial time periods related to these two politicians, i.e., time periods in which there were many positive and negative tweets? How did the "connectedness" of Trump with the entity 'Abortion' evolve during 2016?*

In a nutshell, we make the following contributions:

- We introduce a multi-aspect entity modeling and propose a set of measures for capturing important entity features in a given time period. A sequence of such captures comprises a multi-variate time series in which each point is a multi-aspect description of the entity at a certain time period. We demonstrate the usefulness of our approach through illustrative examples.
- We provide an open source distributed library for computing the proposed measures efficiently.
- We analyze a large Twitter archive (spanning 4 years and containing billions of tweets) and make publicly available the entity- and sentiment- annotations of this archive. This dataset can foster further research in related topics (like event detection, topic evolution, entity recommendation, concept drift).

The rest of this paper is organized as follows: Section 2 provides some background and related works. Section 3 details the multi-aspect entity description and the proposed measures. Section 4 presents a library for the distributed computation of the measures. Section 5 presents the results of a case study. Finally, Section 6 concludes the paper and identifies interesting directions for future research.

2 Background and Related Work

We first discuss the required background and then we describe related works and how they differ from our approach.

2.1 Entity Linking and Sentiment Analysis

Our analysis is based on two different types of annotations applied in the short texts of a social media archive (like a Twitter archive): *entity linking* and *semantic analysis*.

Entity Linking. In our problem, an *entity* is anything with a distinct, separate and meaningful existence that also has a "web identity" expressed through a unique URI (e.g., a Wikipedia/DBpedia URI). This does not only include persons, locations, organizations, etc., but also events (e.g., *US 2016 presidential election*) and concepts (e.g., *Democracy*). Each entity is associated with a unique URI, while several labels/names can be used to refer to this entity. For example, for the entity *Barack Obama* (`https://en.wikipedia.org/wiki/Barack_Obama`), possible names are "Barack Obama", "Obama" and "former President Obama". There is a plethora of tools that automatically extract entities from plain text and link them to knowledge bases like Wikipedia/DBpedia [5, 10, 14] (for a survey on entity linking and resolution, see [9]). In our experiments, we use Yahoo FEL [5] which has been specially designed for linking entities from short texts to Wikipedia.

Sentiment Analysis. Sentiment analysis refers to the problem of assigning a sentiment label (e.g., positive, negative) or sentiment score to a document [15]. We opt for the latest and we use SentiStrength, a robust tool for sentiment strength detection on social web data [21]. SentiStrength assigns both a positive and a negative score (since both types of sentiment can occur simultaneously). The score of a positive sentiment strength score ranges from +1 (not positive) to +5 (extremely positive). Similarly, negative sentiment strength scores range from -1 (not negative) to -5 (extremely negative).

2.2 Related Work

The availability of web-based application programming interfaces (APIs) provided by social media services (like Twitter and Facebook) has led to an "explosion" of techniques, tools and platforms for social media analytics. The work in [4] surveys analytics tools for social media as well as tools for scraping, data cleaning and sentiment analysis on social media data. There is also a plethora of works on exploiting social media for a variety of tasks, like opinion summarization [13], event and rumor detection [3, 16], topic popularity and summarization [2, 23], information diffusion [11], popularity prediction [18], and reputation monitoring [1]. Below, we discuss works related to temporal analysis of topics and entities in social media.

[20] proposes a query-answering framework to allow entity search in social networks by exploiting the underlying social graph and temporal information. [24] studies how to incorporate social attention in the generation of timeline summaries. It proposes capturing social attention for a given topic by learning users' collective interests in the form of word distributions from Twitter. A more recent work on the same topic focuses on how to select a small set of representative tweets to generate a meaningful timeline, which provides enough coverage for a given topical query [23]. [2] performs a spatiotemporal analysis of tweets, investigating the time-evolving properties of the subgraphs formed by the users discussing each topic. The focus is on the network topology formed by follower-following links on Twitter and the geospatial location of the users. [6] introduces a catalogue of metrics for analyzing hashtag-based communication on Twitter,

while [18] tackles the problem of predicting entity popularity on Twitter based on the news cycle. [8] investigates whether semantic relationships between entities can be learned by analyzing microblog posts published on Twitter. The evaluation results showed that co-occurrence based strategies allow for high precision and perform particularly well for relations between persons and events. Our entity-to-entity connectedness scores are also based on entity co-occurrences (more in Section 3).

To our knowledge, our work is the first that models *multi-aspect entity-centric analytics* for social media archives. The proposed measures capture the multi-aspect behavior of an entity in different time periods and can be exploited in a variety of tasks, like entity evolution, event detection, and entity recommendation.

3 Multi-aspect Entity Measures

We propose a multi-aspect description of an entity in terms of its: *popularity* (how much discussion it generates), *attitude* (predominant sentiment), *sentimentality* (magnitude of sentiment), *controversiality* (whether there is a consensus about the sentiment of the entity), *connectedness* to another entity, and *network* (strongly connected entities). All these measures are computed for a given time period (e.g., July 2014, 10-20 June 2013, June-August 2015). Below, we formally introduce these measures by classifying them into: *single-entity measures* and *entity-relation measures*.

First, let C be a collection of short texts (e.g., tweets) covering the time period $T = [t_s, t_e]$ (where t_s, t_e are two different time points with $t_s < t_e$), and let U be the total set of users who posted these texts. Let also E denote a finite set of entities, e.g., all Wikipedia entities.

3.1 Single-Entity Measures

Popularity. Let $e \in E$ be a given entity and $T_i \subseteq T$ a given time period. Let also $C_i \subseteq C$ be the collection of short texts posted during T_i. The popularity of e during T_i equals to the percentage of *texts* mentioning e during that period. Formally:

$$popularity_c(e, T_i) = \frac{|C_{e,i}|}{|C_i|} \tag{1}$$

where $C_{e,i} \subseteq C_i$ denotes the set of texts mentioning e during T_i.

Using the above measure, an entity can be very popular even if it is discussed by a few users but in a large number of texts. A more fine-grained indication of popularity is given by the number of different users discussing the entity. In that case, if $u_c \in U$ denotes the user who posted the text c, the popularity of an entity $e \in E$ during T_i can be defined as the percentage of different *users* discussing e during that period, i.e.:

$$popularity_u(e, T_i) = \frac{|\cup_{c \in C_{e,i}} u_c|}{|\cup_{c \in C_i} u_c|} \tag{2}$$

We can now combine both aspects (percentage of *texts* and *users*) in one popularity score using the following formula:

$$popularity_{c,u}(e, T_i) = popularity_c(e, T_i) \cdot popularity_u(e, T_i) \tag{3}$$

An entity has now a high popularity score if it is discussed in many tweets and by many different users.

Attitude and Sentimentality. We use two measures (proposed in [12] for the case of questions and answers) for capturing a text's *attitude* (predominant sentiment) and *sentimentality* (magnitude of sentiment). First, for a text $c \in C$, let $s_c^+ \in [1, 5]$ be the text's positive sentiment score and $s_c^- \in [-5, -1]$ be the text's negative sentiment score (according to SentiStrength, c.f. Section 2.1). The attitude of a text c is given by $\phi_c = s_c^+ + s_c^-$ (i.e., $\phi_c \in [-4, 4]$) and its sentimentality by $\psi_c = s_c^+ - s_c^- - 2$ (i.e., $\psi_c \in [0, 8]$).

We now define the *attitude* of an entity e in a time period T_i as the average attitude of texts mentioning e during T_i. Formally:

$$attitude(e, T_i) = \frac{\sum_{c \in C_{e,i}} \phi_c}{|C_{e,i}|} \tag{4}$$

Likewise, the *sentimentality* of an entity e in a time period T_i is defined as the average sentimentality of texts mentioning e during T_i:

$$sentimentality(e, T_i) = \frac{\sum_{c \in C_{e,i}} \psi_c}{|C_{e,i}|} \tag{5}$$

Controversiality. An entity e can be considered controversial in a time period T_i if it is mentioned in both many positive and many negative texts. First, let $C_{e,i}^+$ be the set of texts mentioning e during T_i with strong positive attitude, i.e., $C_{e,i}^+ = \{c \in C_{e,i} \mid \phi_c \geq \delta\}$, where $\delta \in [0, 4]$ is a strong attitude threshold (e.g., $\delta = 2.0$). Likewise, let $C_{e,i}^-$ be those with strong negative attitude, i.e., $C_{e,i}^- = \{c \in C_{e,i} \mid \phi_c \leq -\delta\}$. We now consider the following formula for entity *controversiality*:

$$controversiality(e, T_i) = \frac{|C_{e,i}^+| + |C_{e,i}^-|}{|C_{e,i}|} \cdot \frac{min(|C_{e,i}^+|,\ |C_{e,i}^-|)}{max(|C_{e,i}^+|,\ |C_{e,i}^-|)} \tag{6}$$

Intuitively, a value close to 1 means that the probability of the entity being "controversial" is high since there is a big percentage of texts with strong attitude (first part of the formula) and also there are both many texts with strong positive attitude and many texts with strong negative attitude (second part of the formula).

3.2 Entity-Relation Measures

Entity-to-Entity Connectedness. We define a *direct-connectedness* score between an entity $e \in E$ and another entity $e' \in E$ in a time period T_i, as the

number of texts in which e and e' co-occur within T_i. Formally:

$$\textit{direct-connectedness}(e, e', T_i) = \frac{|C_{e,i} \cap C_{e',i}|}{|C_{e,i}|} \tag{7}$$

Notice that the relation is not symmetric. We consider that if an entity e_1 is strongly connected with an entity e_2, this does not mean that e_2 is also strongly connected with e_1. For example, consider that e_1 is mentioned in only 100 texts, e_2 in 1M texts, while 90 texts mention both entities. We notice that e_2 seems to be very important for e_1, since it exists in 90/100 of e_1's texts. On the contrary, e_1 seems not to be important for e_2, since it exists in only 90/1M of its texts.

Two entities may not co-occur in texts, but they may share many common co-occurred entities. For example, both *Barack Obama* and *Donald Trump* may co-occur with entities like *White House*, *US Election* and *Hillary Clinton*. For an input entity $e \in E$ and another entity $e' \in E$, we define an *indirect-connectedness* score which considers the number of *common entities* with which e and e' co-occur in a time period T_i:

$$\textit{indirect-connectedness}(e, e', T_i) = \frac{|(\cup_{c \in C_{e,i}} E_c) \cap (\cup_{c \in C_{e',i}} E_c)|}{|(\cup_{c \in C_{e,i}} E_c)|} \tag{8}$$

where $E_c \subseteq E$ is the entities mentioned in text c. Also in this case, the relation between the two entities is not symmetric.

Entity k-Network. This measure targets at finding a list of entities strongly connected to the query entity in a given time period T_i. First, we define a connectedness score between an entity $e \in E$ and a set of entities $E' \subseteq E$ within T_i, as the average direct-connectedness score of the entities in E'. Formally:

$$\textit{connectedness}(e, E', T_i) = \frac{\sum_{e' \in E'} \textit{direct-connectedness}(e, e', T_i)}{|E'|} \tag{9}$$

The k-Network of an entity e during T_i is the set of k entities $E' \subseteq E$ with the highest average connectedness score. Namely:

$$k\text{-}Network(e, T_i) = \underset{E' \subseteq E,\ |E'|=k}{\operatorname{argmax}} \textit{connectedness}(e, E', T_i) \tag{10}$$

In simple terms, the k-Network of an entity e consists of the k entities with the highest *direct-connectedness* scores.

3.3 Discussion

The above presented measures capture the multi-aspect behavior of a given entity at a certain time period. In the long run, a multi-variate time series is formed where each point represents the multi-aspect description of the entity at a certain period in time.

An important characteristic of our approach is that we can support both entity-specific queries referring to a single entity and cross-entity queries involving more than one entities (e.g., a category of entities). This is achieved through

the *entity linking* process in which entities are extracted from the texts and are linked to knowledge bases like Wikipedia/DBpedia. In that way, we can collect a variety of properties for the entities extracted from our archive. This enables us to aggregate information and capture the behavior of sets of entities. For example, by accessing DBpedia, we can collect a list of German politicians, derive their popularity and then compare it with that of another set of entities.

Although the proposed analysis approach is generic and can be applied over different types of social media archives, it is clear that the quality of the generated data depends on the quality of the input data. Twitter, for example, provides 1% random sample, which though is subject to bias, fake news and possibly other adversarial attacks. In our case study (detailed in Section 5), although we remove spam, we do not take similar actions to deal with bias and other data peculiarities. This also means that high profile entities might occupy a big volume in the archive, whereas long-tail entities might be underrepresented or not represented at all. Except for the quality of the original data, the different preprocessing steps (spam removal, entity linking, sentiment analysis) are also prone to errors. This means that, especially for small archives, the data produced by the proposed measures are also prone to errors. For instance, regarding the entity linking task, selecting a very low threshold for the confidence score of the extracted entities can result in many false annotations, which in turn can affect the quality and reliability of the produced time-series.

4 Library for Computing the Measures

For computing the measures, we provide an Apache Spark library. Apache Spark[4] is a cluster-computing framework for large-scale data processing. The library contains functions for computing the proposed measures for a given entity and over a specific time period. It operates over an annotated (with entities and sentiments) dataset split per year-month (the dataset should be in a simple CSV format). The library is available as open source[5].

The time for computing the measures highly depends on the dataset volume, the used computing infrastructure as well as the available resources and the load of the cluster at the analysis time. The Hadoop cluster used in our experiments for analyzing a large Twitter archive of more than 1 billion tweets consisted of 25 computer nodes with a total of 268 CPU cores and 2,688 GB RAM (more about the dataset in the next section). Indicatively, the time for computing each of the measures was on average less than a minute (without using any index, apart from the monthly-wise split of the dataset).

5 Case Study: Entity Analytics on a Twitter Archive

In this section, we first describe the results of the analysis and annotation of a large Twitter archive. Then, we present examples of case studies illustrating the insights gained from the proposed measures.

[4] `http://spark.apache.org/`
[5] `https://github.com/iosifidisvasileios/Large-Scale-Entity-Analysis`

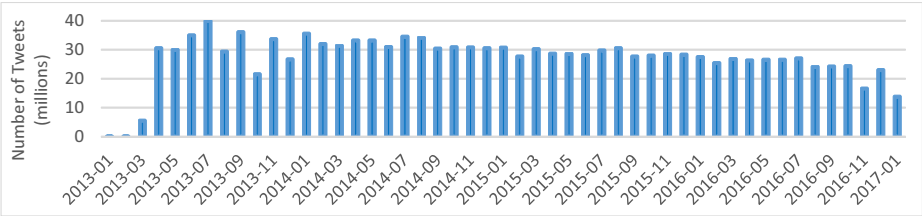


Fig. 1: Number of tweets per month.

5.1 Annotating a Large Twitter Archive

We analyzed a large Twitter archive spanning 4 years (January 2014 - January 2017) and containing more than 6 billion tweets. The tweets were collected through the Twitter streaming API. Our analysis comprised the following steps: i) filtering (filtering out re-tweets, keeping only English tweets), ii) spam removal, iii) entity linking, and iv) sentiment analysis. The filtering step reduced the number of tweets to about 1.5 billion tweets (specifically, to 1,486,473,038 tweets). For removing the spam tweets, we trained a Multinomial Naive Bayes (MNB) classifier over the HSpam dataset [19]. This removed about 150 million tweets. The final dataset consists of 1,335,324,321 tweets from 110,548,539 users. Figure 1 shows the number of tweets per month on the final dataset.

For the *entity linking* task, we used Yahoo FEL [5] with a confidence threshold score of -3. Totally, 1,390,286 distinct entities were extracted from the tweets collection. On average, each tweet contains about 1 entity (specifically, 0.95), while FEL returned no entity for about 550 million tweets. For each extracted entity, we also store the confidence score provided by FEL. Thereby, data consumers can select suitable confidence ranges to consider, depending on the specific requirements with respect to precision and recall. For *sentiment analysis*, we used SentiStrength [21]. The average sentimentality of all tweets is 0.92, the average attitude 0.2, while 622,230,607 tweets have no sentiment (-1 negative sentiment and 1 positive sentiment). Table 1 shows the number of tweets per attitude value.

The annotated dataset is publicly available in CSV format[6]. We make the dataset available so anyone interested can use it together with the library (described in Section 4) to extract the measures for any entity at the desired level of temporal granularity.

Table 1: Number of tweets per attitude value.

Attitude:	-4	-3	-2	-1	0
Number of tweets:	2,234,887	34,666,708	68,812,370	104,628,022	670,484,267

Attitude:	1	2	3	4
Number of tweets:	301,635,430	138,197,637	13,610,492	1,054,508

[6] `http://l3s.de/~iosifidis/tpdl2017/`. For each tweet the dataset includes the following information: ID, user (encrypted), post date, extracted entities, positive and negative sentiment values. The text of the tweets is not provided for copyright purposes.

5.2 Case Studies

Entity Popularity. Figure 2 (left) shows the popularity of *Alexis Tsipras* (Greek prime minister) within 2015. We notice that his popularity highly increased in July. Indeed, in July 2015 the Greek bailout referendum was held following the bank holiday and capital controls of June 2015. This event highly increased the popularity of the Greek prime minister. Moreover, by comparing the trend of the two different popularity scores (Formulas 1 and 2), we notice that, during June and July 2015, the percentage of different users discussing about *Alexis Tsipras* increased in bigger degree compared to the percentage of tweets, implying that more people were engaged in the discussion.

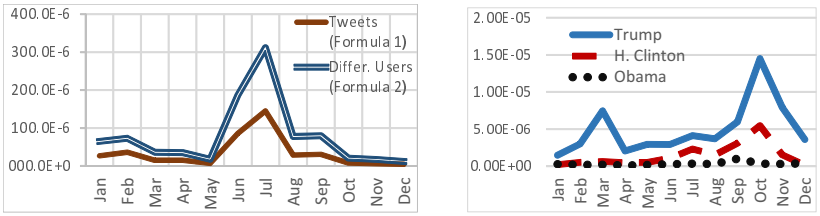


Fig. 2: Popularity of *"Alexis Tsipras"* in 2015 (left); Popularity of *"Donald Trump"*, *"Hillary Clinton"* and *"Barack Obama"* in 2016 (right).

Likewise, we can compare the popularity of multiple entities within the same time period. For example, Figure 2 (right) shows the popularity of *Donald Trump*, *Hillary Clinton* and *Barack Obama* within 2016 (according to Formula 3). We notice that *Donald Trump* is much more popular in all months. We also notice that, in October 2016 the popularity of *Donald Trump* and *Hillary Clinton* highly increased compared to the other months. This is an indicator of possible important events related to these two entities in October 2016 (indeed, two presidential general election debates took place in that period).

Entity Attitude and Sentimentality. Figure 3 (left and middle) depicts the attitude and sentimentality of *Donald Trump* and *Hillary Clinton* within 2016. We notice that both entities had constantly a negative attitude, however that of *Hillary Clinton* was worse in almost all months. Moreover, we notice that *Hillary Clinton*'s attitude highly decreased in May 2016 (possibly, for example, due to a report issued by the State Department related to Clinton's use of private email). Regarding sentimentality, we notice that for the majority of months the tweets mentioning *Donald Trump* are a bit more sentimental than those mentioning *Hillary Clinton*. In general, we notice that the values of both attitude and sentimentality are relatively small and close to zero. This is due to the very big number of tweets with no sentiment (almost half of the tweets).

Entity Controversiality. Figure 3 (right) shows the controversiality of *Donald Trump* and *Hillary Clinton* within 2016 (using $\delta = 2.0$). We notice that *Donald Trump* induces more controversial discussions in Twitter than *Hillary Clinton*, while February was his most "controversial" month, probably because of his references to some debatable topics (like abortion) during his campaign trail.

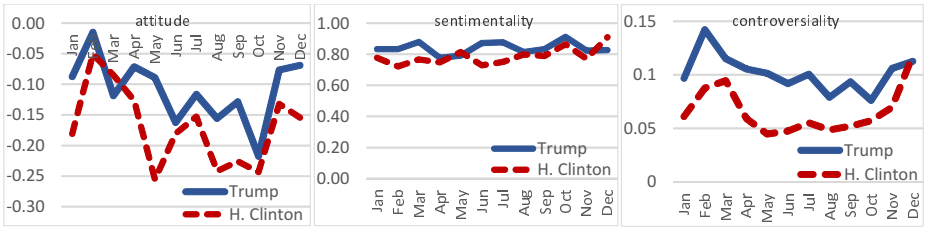
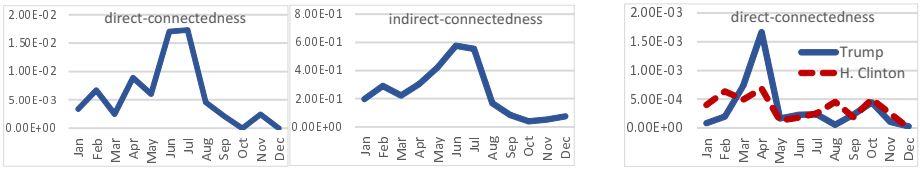


Fig. 3: Attitude (left), sentimentality (middle) and controversiality (right) of *"Donald Trump"* and *"Hillary Clinton"* in 2016.

It is interesting also that *Hillary Clinton*'s controversiality has an exponential increment from September to December 2016.

Entity-to-Entity Connectedness. Figure 4 (a) depicts the connectedness of *Alexis Tsipras* with the concept *Greek withdrawal from the eurozone* within 2015. We notice that these two entities are highly connected in June and July, while after August, their connectedness is very close to zero. Indeed, important events related to Greece's debt crisis took place in June and July 2015, including the bank holiday, the capital controls and the Greek bailout referendum. Likewise, Figure 4 (b) shows the connectedness of both *Donald Trump* and *Hillary Clinton* with the concept *Abortion* in 2016. Here we notice that the connectedness is almost constant for *Hillary Clinton*, while for *Donald Trump*, there is a very large increment in March and April.



(a) (b)

Fig. 4: (a) Connectedness of *"Alexis Tsipras"* with *"Greek withdrawal from the eurozone"* in 2015 (Formulas 7 and 8); (b) Connectedness of *"Donald Trump"* and *"Hillary Clinton"* with *"Abortion"* in 2016 (Formula 7).

Entity k-Network. Figure 5 shows the 10-Network of *Alexis Tsipras* in three different time periods (April, July and October, 2015). We notice that there are three general entities that exist in all time periods (*Greece*, *Athens*, *Reuters*). For April and July, we notice that the 10-Network contains 4 common entities (*Syriza*, *Referendum*, *Greek withdrawal from the eurozone*, and *Yanis Varoufakis*), while for July and October, *Austerity* is the only common entity (probably related to the approval of strict measures required by the creditors). For April, the 10-Network contains three entities related to Russia (due to Tsipra's visit in Moscow to meet Russian president Vladimir Putin), while for October, it contains two entities related to European migrant crisis (probably due to Tsipra's visit in Lesvos island).

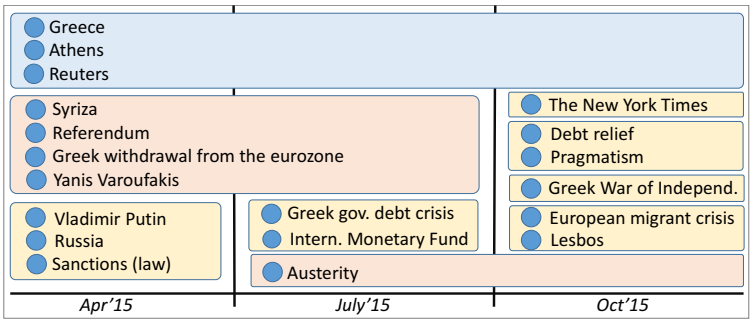


Fig. 5: 10-Network of *Alexis Tsipras* in April, July and October 2015.

6 Conclusion

We have proposed an entity-centric and multi-aspect approach to analyze social media archives, and we defined measures that allow studying how entities are reflected in social media and how entity-related information evolves over time. We believe that the proposed analysis approach is the first step towards more advanced and meaningful exploration of social media archives, while it can facilitate research in a variety of fields, such as information extraction, sociology, and digital humanities.

As part of our future work, we plan to exploit the rich amount of generated data for *prediction* of entity-related features. In particular, given an entity, our focus will be on how we can predict future values of the proposed measures (e.g., popularity or attitude in a given horizon). We also intend to study approaches on *understanding* and *representing* the dynamics of such evolving entity-related information, for instance, as done in [17] for the case of RDF datasets.

Acknowledgements. The work was partially funded by the European Commission for the ERC Advanced Grant ALEXANDRIA under grant No. 339233.

References

1. E. Amigó, J. Carrillo-de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, E. Meij, M. de Rijke, and D. Spina. Overview of RepLab 2014: author profiling and reputation dimensions for online reputation management. In *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 2014.
2. S. Ardon, A. Bagchi, A. Mahanti, A. Ruhela, A. Seth, R. M. Tripathy, and S. Triukose. Spatio-temporal analysis of topic popularity in Twitter. *arXiv preprint arXiv:1111.2904*, 2011.
3. F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. *Computational Intelligence*, 31(1), 2015.
4. B. Batrinca and P. C. Treleaven. Social media analytics: a survey of techniques, tools and platforms. *AI & SOCIETY*, 30(1), 2015.
5. R. Blanco, G. Ottaviano, and E. Meij. Fast and space-efficient entity linking for queries. In *WSDM*, 2015.
6. A. Bruns and S. Stieglitz. Towards more systematic Twitter analysis: metrics for tweeting activities. *Internat. Journal of Social Research Methodology*, 16(2), 2013.

7. A. Bruns and K. Weller. Twitter as a first draft of the present: and the challenges of preserving it for the future. In *8th ACM Conference on Web Science*, 2016.
8. I. Celik, F. Abel, and G.-J. Houben. Learning semantic relationships between entities in Twitter. In *International Conference on Web Engineering*, 2011.
9. V. Christophides, V. Efthymiou, and K. Stefanidis. *Entity Resolution in the Web of Data*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, 2015.
10. P. Ferragina and U. Scaiella. Tagme: on-the-fly annotation of short text fragments (by Wikipedia entities). In *CIKM*, 2010.
11. A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2), 2013.
12. O. Kucuktunc, B. B. Cambazoglu, I. Weber, and H. Ferhatosmanoglu. A large-scale sentiment analysis for Yahoo! answers. In *WSDM*, 2012.
13. X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang. Entity-centric topic-oriented opinion summarization in Twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.
14. A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2, 2014.
15. B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2), 2008.
16. V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.
17. Y. Roussakis, I. Chrysakis, K. Stefanidis, G. Flouris, and Y. Stavrakas. A Flexible Framework for Understanding the Dynamics of Evolving RDF Datasets. In *ISWC*, 2015.
18. P. Saleiro and C. Soares. Learning from the News: Predicting Entity Popularity on Twitter. In *International Symposium on Intelligent Data Analysis*, 2016.
19. S. Sedhai and A. Sun. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.
20. K. Stefanidis and G. Koloniari. Enabling Social Search in Time through Graphs. In *Web-KR@CIKM*, 2014.
21. M. Thelwall, K. Buckley, and G. Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.
22. G. Weikum, M. Spaniol, N. Ntarmos, P. Triantafillou, A. Benczúr, S. Kirkpatrick, P. Rigaux, and M. Williamson. Longitudinal Analytics on Web Archive Data: It's About Time! In *CIDR*, 2011.
23. J.-g. Yao, F. Fan, W. X. Zhao, X. Wan, E. Chang, and J. Xiao. Tweet timeline generation with determinantal point processes. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016.
24. X. W. Zhao, Y. Guo, R. Yan, Y. He, and X. Li. Timeline generation with social attention. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013.
25. M. Zimmer. The Twitter Archive at the Library of Congress: Challenges for information practice and information policy. *First Monday*, 20(7), 2015.