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

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

fafalios@L3S.de

[1] L3S Research Center
University of Hannover, Germany

[2] Faculty of Natural Sciences,
University of Tampere, Finland

Motivation

- Social media archives serve as **important historical information sources** and are of **immense value for future generations**
- Initiatives have started **collecting** and **preserving** such user-generated data (like the Twitter Archive at the Library of Congress)

- Absence of **meaningful access** and **analysis** methods
  - [Bruns and Weller, ACM Conference on Web Science, 2016]
- Analysts want to see, compare, and understand trends about **entities**
  - Thus calling for **entity-level** analytics over the archived data!

Motivating Questions

- How did the **popularity** of an entity evolve in a specific time period? Were there any “outlier” periods? What **other entities** were discussed in social media together with the query entity during these periods?

- What was the **predominant sentiment** about an entity in a specific time period and how did it evolve over time? Were there any “**controversial**” time periods related to that entity?

- How did the “**connectedness**" of an entity with another entity evolve during a time period? What may have affected an increase in their connectedness?

Approach overview

- Apply **entity linking** and **sentiment analysis** on the (short) texts of a social media archive
  - **Entity Linking**: extract named entities from plain text and link them to a knowledge base (e.g., Wikipedia/DBpedia)
  - **Sentiment Analysis**: assign a sentiment label (e.g., positive/negative) or sentiment score to a text
- Compute **measures** that characterize different aspects of the entities in different time periods
  - *Entity Popularity*
  - *Entity Attitude (predominant sentiment)*
  - *Sentimentality (strength of sentiment)*
  - *Controversiality (many positive and many negative sentiments)*
  - *Entity-to-Entity Connectedness*
  - *Entity k-Network*

Contributions

- We introduce a **multi-aspect entity modeling** and a **set of measures** for capturing important **entity features** in a specific time-period
  - A sequence of such captures comprises a time series
  - We demonstrate the usefulness of the proposed approach through illustrative examples
- We provide an **open source library** (apache spark) for the efficient computation of the introduced measures
- We analyze a **large Twitter archive** (spanning **4 years** and containing **billions** of tweets)
  - We make **publicly available** the entity and sentiment annotations of this archive

Outline

- Background
  - Entity linking
  - Sentiment Analysis
  - Related Works
- Multi-aspect Entity Measures
  - Single-entity measures
  - Entity-relation measures
  - Library for computing the measures
- Case Study
  - Entity analytics on a large Twitter archive
- Conclusion and Future Work

Background

- Entity Linking
- Sentiment Analysis
- Related Works

Background

- **Entity Linking**: extract named entities from plain text and link them to a reference knowledge base (e.g., Wikipedia/DBpedia)

"Obama's speech during his Houston visit was very good!"

(confidence: 92%) → https://en.wikipedia.org/wiki/**Barack_Obama**

(confidence: 86%) → https://en.wikipedia.org/wiki/**Houston**

Yahoo FEL

- Reference knowledge base: Wikipedia
- Very lightweight and efficient, good accuracy

Background

- **Sentiment Analysis**: assign a sentiment label (e.g., positive/negative) or sentiment score to a text

"Obama's speech during his Houston visit was very good!" → Positive: 80%, Negative: 0%

"I love dogs but I hate cats" → Positive: 100%, Negative: 100%

SentiStrength

- Robust tool for sentiment strength detection
- It assigns both a positive and a negative score
  - Positive score ranges from +1 (not positive) to +5 (extremely positive)
  - Negative score ranges from -1 (not negative) to -5 (extremely negative)

Related Works

- **Tools** for analytics, cleaning and sentiment analysis on social media data [survey by Batrinca and Treleaven, 2015]
- **Exploiting** Social Media for:
  - Event detection [Atefeh and Khreich, 2015]
  - Topic summarization [Yao et al., 2016]
  - Information diffusion [Guille at al., 2013]
  - Reputation monitoring [Amigo et al., 2014]
- **Temporal analysis** of topics and entities in social media:
  - Social search in time [Stefanidis and Koloniari, 2014]
  - Timeline summaries [Zhao et al., 2013]
  - Spatiotemporal analysis of topic popularity [Ardon et al., 2011]
  - Popularity detection [Saleiro and Soares, 2016]

Multi-aspect Entity Measures

- Single-entity measures
- Entity-relation measures
- Library for computing the measures

Measures

- Single-entity measures:
  - Popularity
  - Attitude
  - Sentimentality
  - Controversiality
- Entity-relation measures:
  - Entity-to-Entity Connectedness
  - Entity k-Network

Computed for a specific time period of any granularity

- e.g., July 2014, 10-20 May 2013, …

Single-Entity Measures

➢ Popularity

- Percentage of **<u>posts</u>** mentioning the query entity during a given time period

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

- Percentage of **<u>different users</u>** mentioning the query entity in a given time period

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

- Combination:

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

Single-Entity Measures

➢ Attitude

- **<u>predominant sentiment</u>** of posts mentioning the query entity
- Attitude for single text = positive score + negative score.
  - Example: (+4) + (-2) = +2

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

➢ Sentimentality

- **<u>magnitude of sentiment</u>** of posts mentioning the query entity
- Sentimentality for single text = |positive score| + |negative score|
  - Example: |+4| + |-2| = +6

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

Single-Entity Measures

- **Controversiality**
  - Big number of posts with strong positive attitude **AND** big number of posts with strong negative attitude

$$controversiality(e, T_i) = \underbrace{\frac{|C^+_{e,i}| + |C^-_{e,i}|}{|C_{e,i}|}}_{\text{Percentage of posts with strong attitude}} \cdot \underbrace{\frac{min(|C^+_{e,i}|,\ |C^-_{e,i}|)}{max(|C^+_{e,i}|,\ |C^-_{e,i}|)}}_{\text{Ratio of posts with strong positive attitude and strong negative attitude}}$$

Entity-Relation Measures

- **Entity-to-Entity Connectedness**
  - Direct: co-occurrence in posts

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

  - Indirect: shared co-occurring entities

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

Not symmetric!

Entity-Relation Measures

- **Entity k-Network**
  - Entities strongly connected with the query entity in a given time period

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

Connectedness between an entity an a set of other entities

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

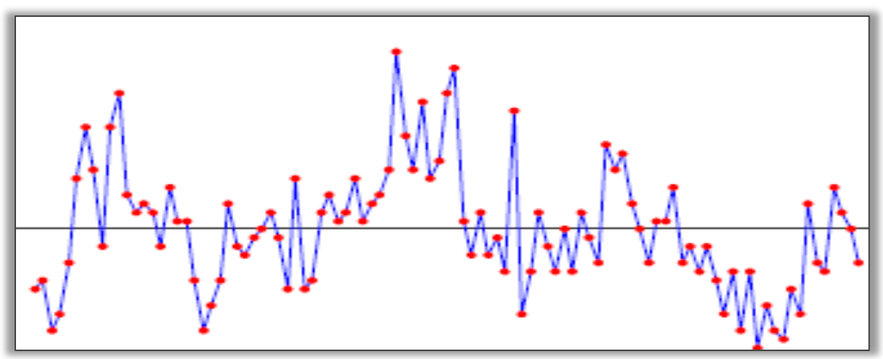
Computing the measures

- Open Source Apache Spark library
- Compute the measure for any given **entity** and **time period**
- Operates over an annotated (with entities and sentiments) dataset split per year-month
- Efficiency highly depends on
  - dataset volume
  - computing infrastructure
  - available resources and load of the cluster

https://github.com/iosifidisvasileios/Large-Scale-Entity-Analysis

Discussion

- Generation of time-series data
  - E.g., month-wise



- Support for both single entities and category of entities (by exploiting the links to the reference knowledge base)
  - E.g., popularity of German politicians

- Quality of derived data depends on quality of input data
  - Bias / Fake / Spam
  - Entity Linking and Sentiment Analysis are prone to errors

Case Study

- Entity analytics on a large Twitter archive

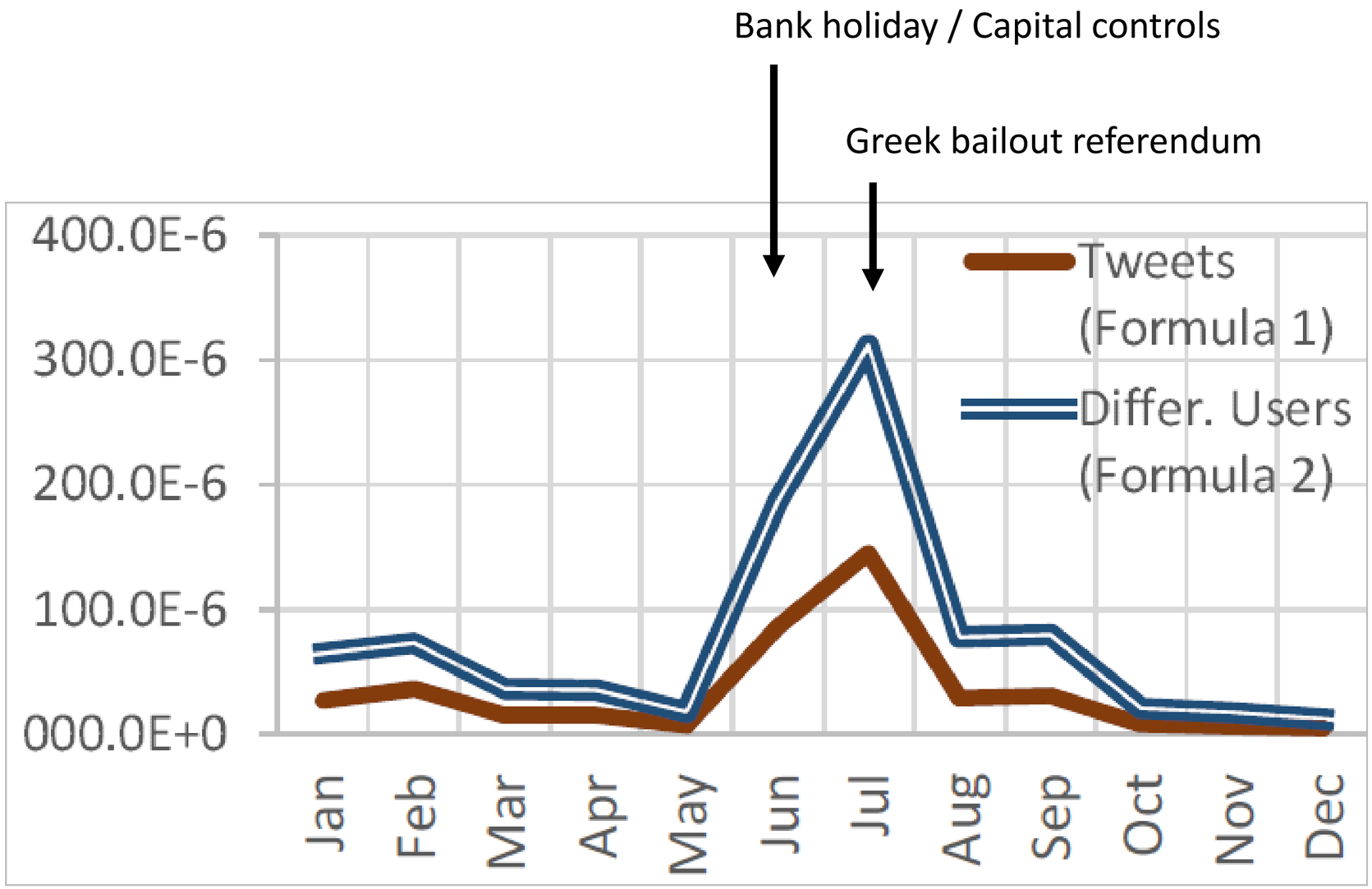
Case Study: Entity Analytics on a Twitter Archive

Dataset:

- Twitter archive spanning 4 years (Jan'14 – Jan'17), containing > 6 billion tweets
- Analysis steps:
  - Filtering (filtering out re-tweets and non-English tweets)
  - Spam removal (by training a MNB classifier using the HSpam dataset)
  - Entity linking (using Yahoo FEL)
  - Sentiment Analysis (using SentiStrength)
- Final dataset:
  - ≈ **1.3 billion** tweets from 110 million users
  - ≈ 1.3 million distinct entities
  - ≈ 700 million tweets with sentiment

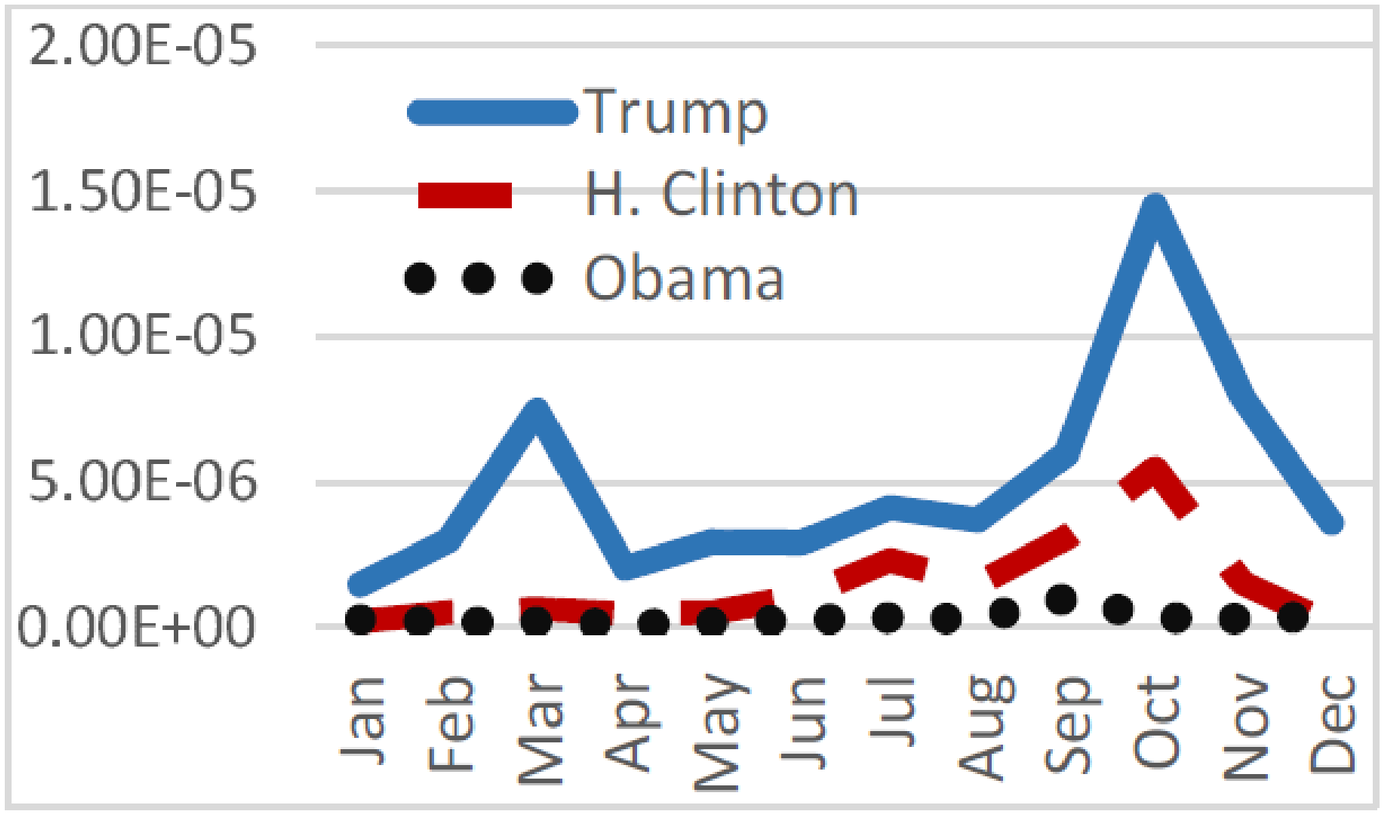
Publicly available in the CSV format: http://l3s.de/~iosifidis/tpdl2017/

Popularity
(Alexis Tsipras in 2015)

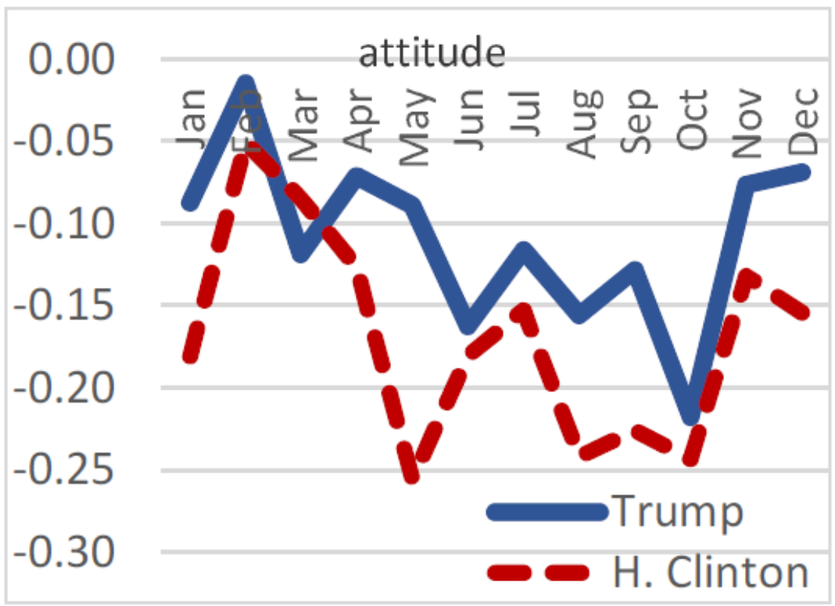


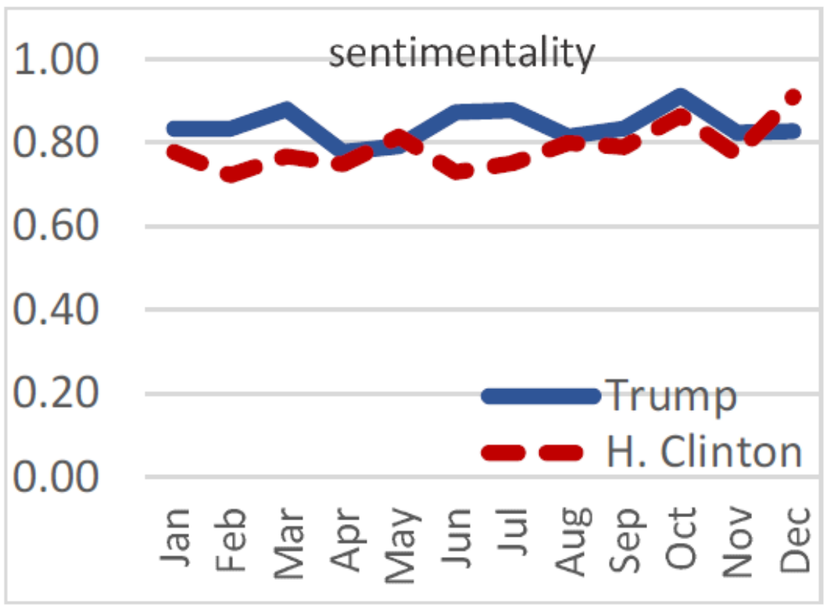
Popularity

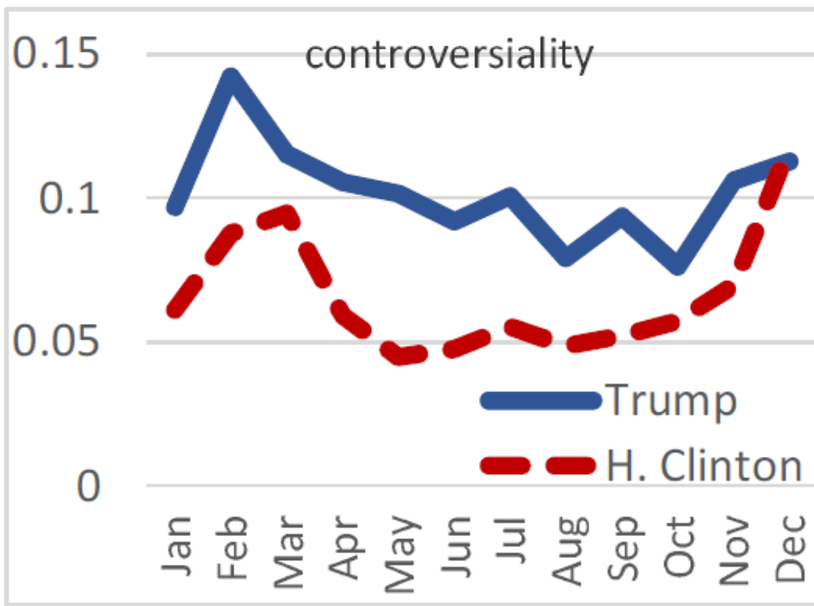
(D. Trump, H. Clinton and B. Obama in 2016)



Attitude, **Sentimentality** and **Controversiality**
(D. Trump and H. Clinton in 2016)

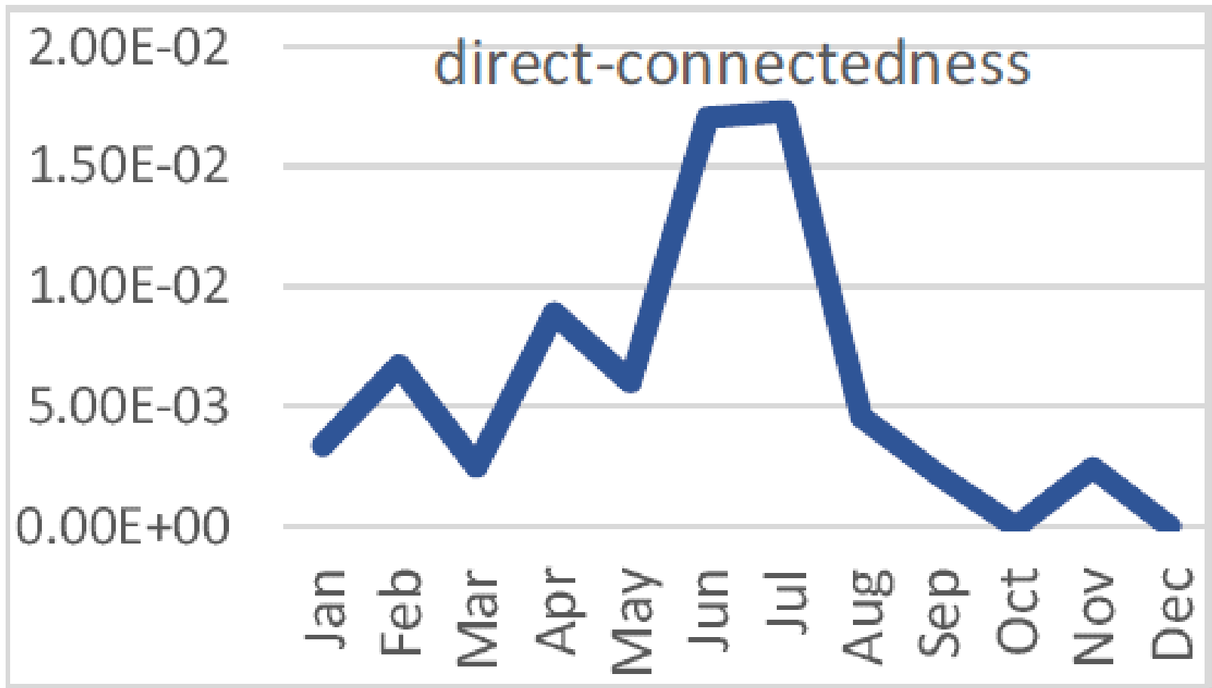


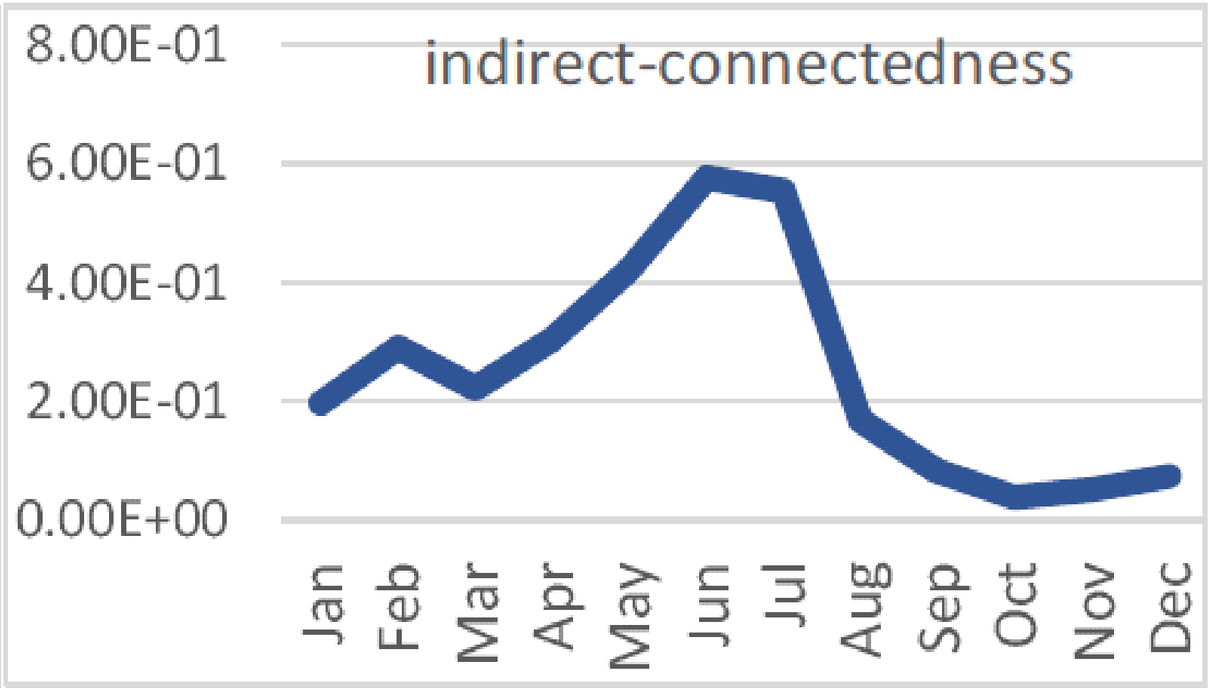




Entity-to-entity connectedness

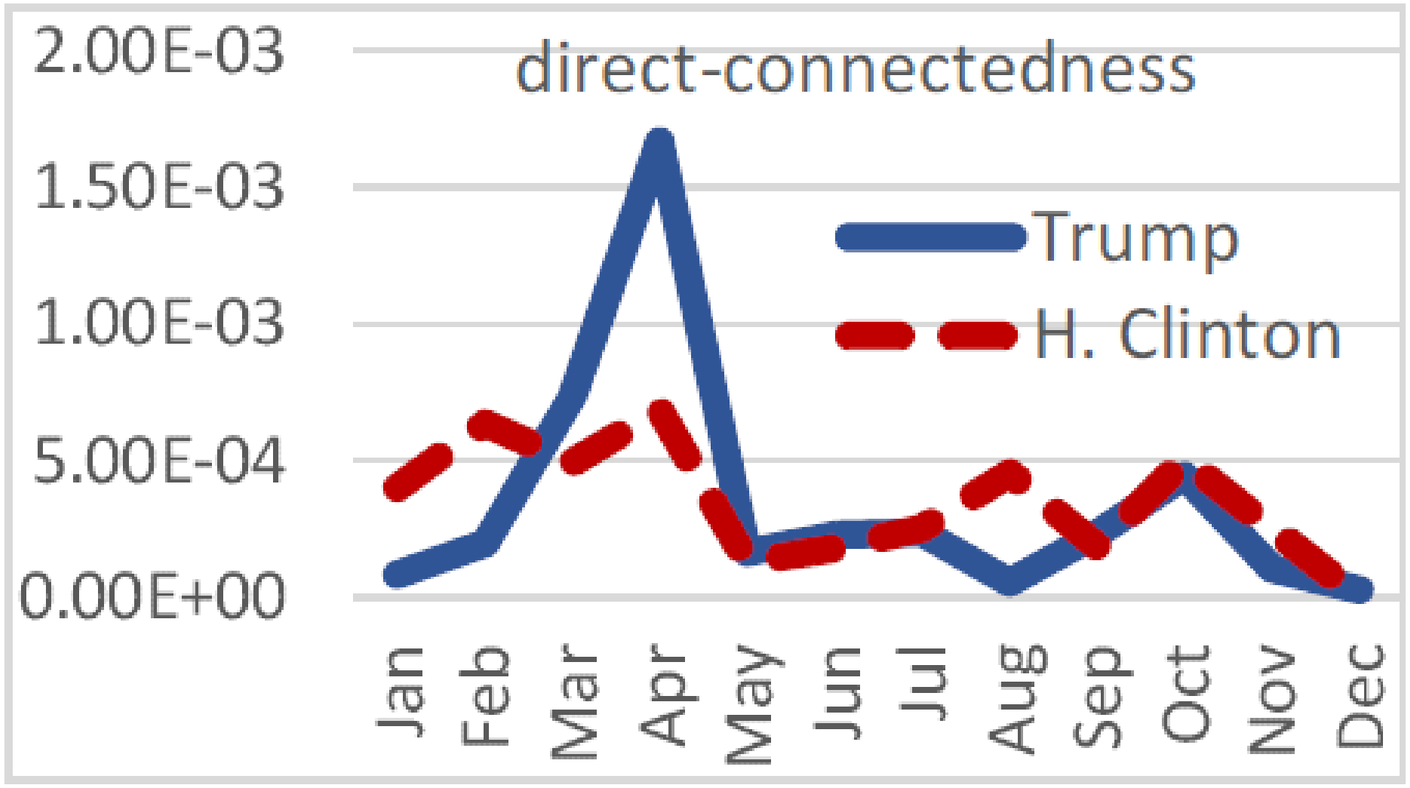
(Alexis Tsipras with *"Greek withdrawal from the Eurozone"* in 2015)





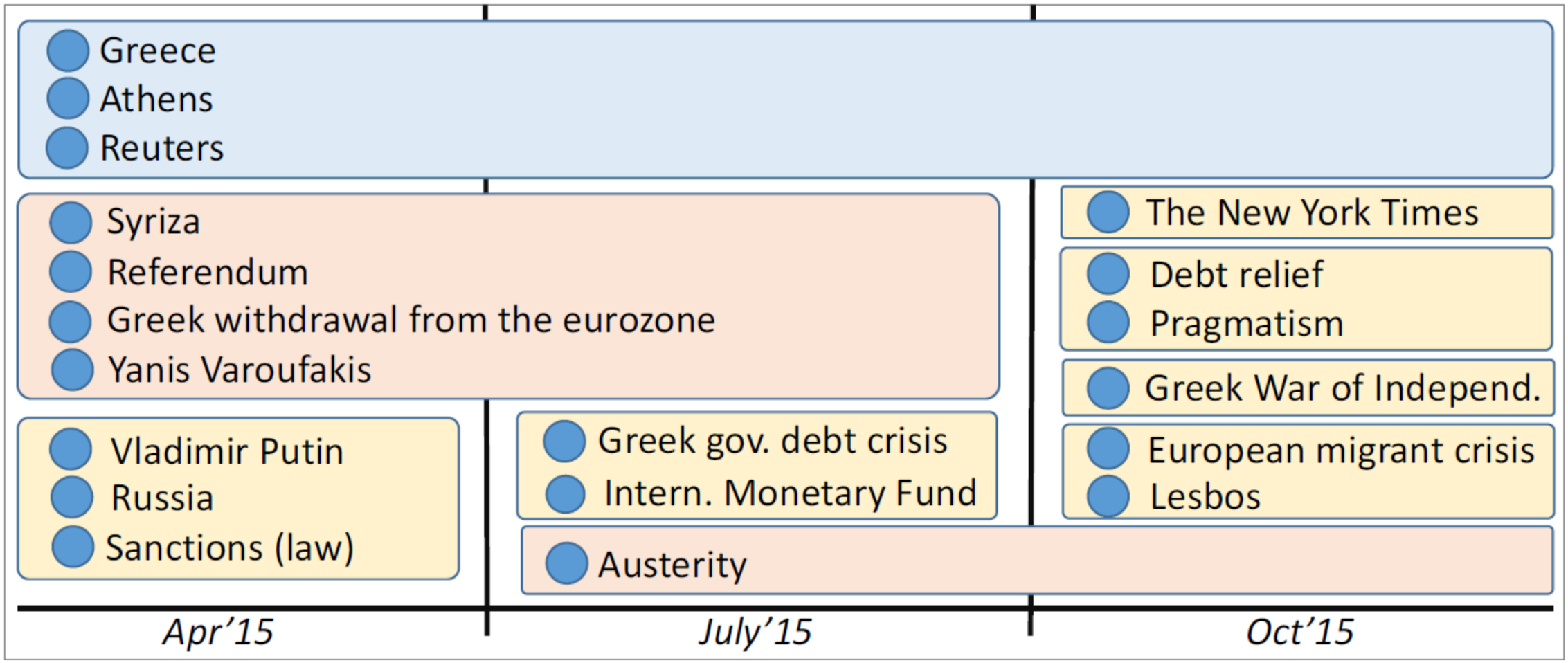
Entity-to-entity connectedness

(D. Trump and H. Clinton with *“Abortion”* in 2015)



Entity *k*-Network

(10-Network of Alexis Tsipras in April, July and October 2015)



Conclusion and Future Work

- **Entity-centric** and **Multi-aspect** approach to analyze social media archives
- **Measures**:
  - Popularity, Attitude, Sentimentality, Controversiality
  - Entity-to-Entity Connectedness, Entity k-Network
- **Facilitate research** in a variety of fields
  - Information extraction, Sociology, Digital humanities, …

- **Future work:**
  - **Prediction** of entity-related features
  - Understanding and representing the **dynamics** of evolving entity-related information

Thank you!
Questions?

http://www.alexandria-project.eu/
(ERC Advanced Grant)