Issue 2 / February 2019

SOBIGDATA RESEARCH INFRASTRUCTURE

Social Mining & Big Data Ecosystem

SoBigData

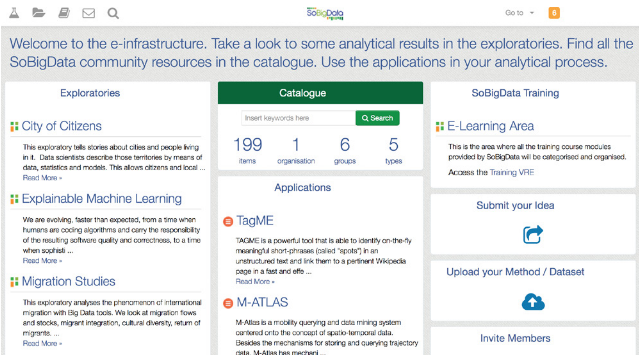
RESEARCH INFRASTRUCTURE

Magazine

Editorial

The H2020 Research Infrastructure SoBigData.eu - Social Mining & Big Data Ecosystem - funded by the European Commission and coordinated by the Italian CNR, has worked since 2015 to integrate national communities and resources and create a vibrant and solid platform for open, ethically minded data science research and innovation.

SoBigData has placed a special focus on fostering novel interdisciplinary research on social challenges underpinned by big data sources and the powerful analytical tools of data mining, machine learning and network science.



Welcome to the e-infrastructure. Take a look to some analytical results in the exploratories. Find all the SoBigData community resources in the catalogue. Use the applications in your analytical process.

Exploratories

City of Citizens

This exploratory tells stories about cities and people living in it. Data scientists describe those territories by means of data, statistics and models. This allows citizens and local ...
Read More »

Explainable Machine Learning

We are evolving, faster than expected, from a time when humans are coding algorithms and carry the responsibility of the resulting software quality and correctness, to a time when sophisti ...
Read More »

Migration Studies

This exploratory analyses the phenomenon of international migration with Big Data tools. We look at migration flows and stocks, migrant integration, cultural diversity, return of migrants. ...
Read More »

Catalogue

Insert keywords here | Search

199	1	6	5
items	organisation	groups	types

Applications

TagME

TAGME is a powerful tool that is able to identify on-the-fly meaningful short-phrases (called "spots") in an unstructured text and link them to a pertinent Wikipedia page in a fast and effe ...
Read More »

M-ATLAS

M-Atlas is a mobility querying and data mining system centered onto the concept of spatio-temporal data. Besides the mechanisms for storing and querying trajectory data, M-Atlas has mechani ...

SoBigData Training

E-Learning Area

This is the area where all the training course modules provided by SoBigData will be categorised and organised.

Access the Training VRE

Submit your Idea

Upload your Method / Dataset

Invite Members

[Continues on pag. 3]

Inside this issue

Content

Editorial

News

Experiences from the TNA calls

SoBigData Datathons

Research Highlights

Editorial

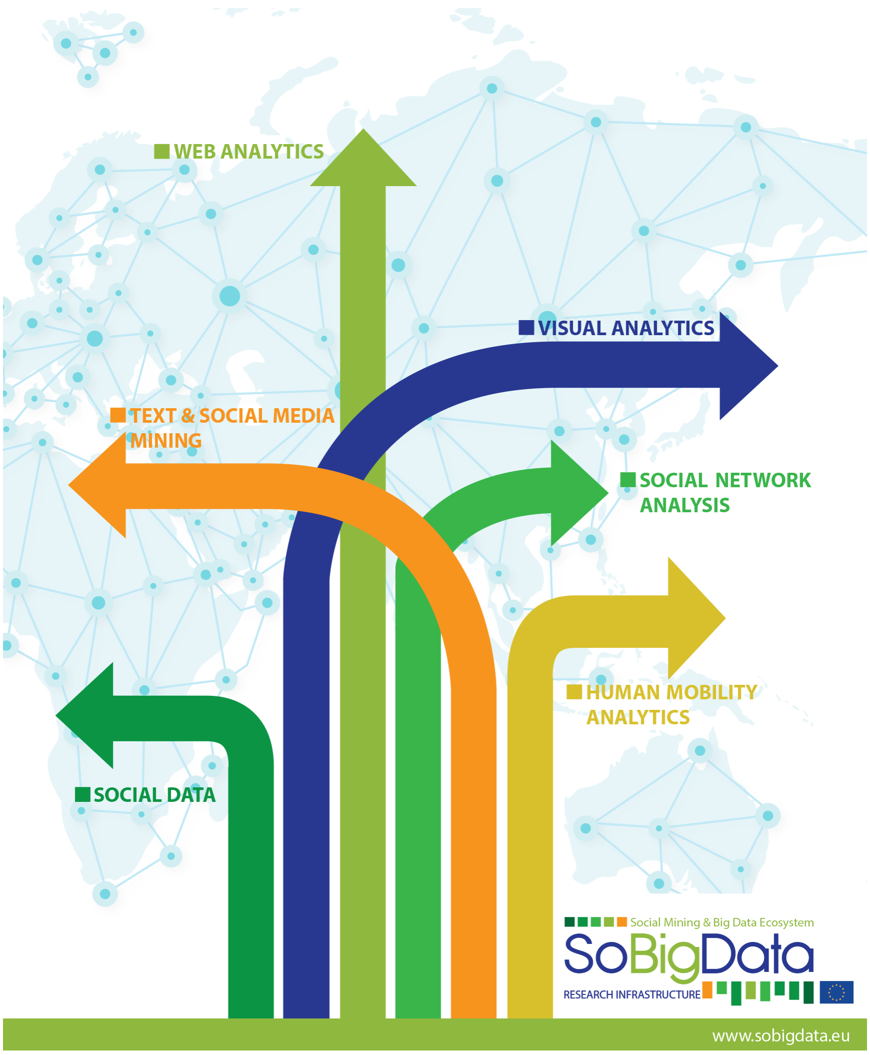
The future of the project, the ambition to broaden the SoBigData community and consolidate his platform and ecosystem, ensuring a sustainable and impactful future.

Dino Pedreschi, Fosca Giannotti, Kalina Bontcheva, Roberto Trasarti

[continued]

After three years since its start, the SoBigData approach revealed successful in delivering a comprehensive e-infrastructure offering access to 180+ social datasets, big data analysis algorithms and courseware, to support data scientists for executing large-scale experiments. The e-infrastructure is used by a wide volume and diversity of stakeholders, including

120 companies and 2500+ registered users, with peaks of million accesses and executions per day. On top of this platform, SoBigData has created five „exploratories“, or vertical research environments focused on broad societal challenges: City of citizens, Societal debates, Well-being and economic performance, Migration studies, and Sport analytics. The exploratories are the places where concrete, substantive multi-disciplinary research is carried on. We funded the on-site visits by 35 data scientists from outside the consortium that were supported and hosted so far to perform their projects, thanks to the Transnational Access action of SoBigData. The activity around the exploratories has produced 200+ high-profile publications powered by big data experiments executed by the SoBigData e-infrastructure, and the associated training and innovation actions have produced courses for 700+ students and 120 R&D pilot projects with companies (53 SMEs, 41 large companies, 26 public institutions).



We are currently working to consolidate the construction of SoBigData and expand its impact. We are in the process of applying for the continuation of SoBigData, with the ambition of becoming an "advanced community". Leveraging the results achieved so far, we want to broaden our community and consolidate our platform and ecosystem, ensuring a sustainable and impactful future for SoBigData. The relevant call is INFRAIA-01-2018-2019, Integrating Activities for Advanced Communities. We will be applying to extend the construction of SoBigData for four further years, 2019-2023. One of the new goals of the consolidation phase will be to create an association of stakeholders interested to become active part in the SoBigData community, and to interface SoBigData towards the European Open Science Cloud as it becomes operational.

We are also working to strengthen SoBigData's ties with two large-scale initiatives around human-centric artificial intelligence that just started, which also have several SoBigData partners in their consortium: the AI-4EU H2020 project, which is mobilising the whole European AI ecosystem and unites 79 partners in 21 countries across Europe (http://ai-4eu.org); and the Humane AI H2020 Flagship proposal coordination action, a pan-European initiative to design and deploy AI systems that enhance human capabilities and empower people – both as individuals and society as a whole – to develop AI that extends rather than replaces human intelligence.

The spectacular advances of artificial intelligence tools, such as automated language understanding and translation, image recognition and robotic vision, automated game players capable of outperforming humans, are essentially successes of data science, explained by the synergic effect of rich data, data-driven learning models, and computing power. Therefore, there is a close link between AI, big data and data science at the root of the current hype and high expectations around AI. A key enabler to the human-centric vision of AI is "explainable AI", or "interpretable machine learning", i.e., the ability to make the decision behavior of ML "black box" models understandable in human terms, so to interface the AI tools with the human stakeholders, data scientists, AI developers, domain experts, or simply citizens.

The SoBigData community has unique skills and resources to build tools for explainable AI, and has recently started a dedicated exploratory that provides a playground for experimenting with such explanation tools, devising new ones, and interfacing with communities of users to validate the effectiveness of the various approaches in real applications. We are happy to see a huge interest along these lines, both from AI and data science researchers and a diversity of companies, which increasingly recognize the importance of validating the logic of the AI components of their services and products for ensuring safety as well as for gaining the trust of their customers.

If you find SoBigData interesting, and believe that it may help you in your research or innovation work, please do hesitate to reach out and express your support to our initiative. We strive to contribute to develop a strategy for data science and artificial intelligence research and innovation at European, national and regional level, and your interest and support adds value and credibility to SoBigData, increasing its chances of success.

Fosca Giannotti, *CNR, IT,*
SoBigData.eu coordinator

Kalina Bontcheva, *University of Sheffield, UK,*
SoBigData.eu deputy coordinator

Dino Pedreschi, *University of Pisa, IT,*
SoBigData.eu deputy coordinator

Roberto Trasarti, *CNR, IT,*
SoBigData.eu project manager

BECOME ASSOCIATE

If you are interested in becoming a SoBigData associate member, follow the link below.

Becoming a SoBigData Associate Member is **completely free** and includes a set of services provided by SoBigData.

Get to know the data scientist community: your organization will be listed and your logo will be shown in our website.

Stay informed: you will receive the SoBigData magazine about the main events related to the research infrastructure, and important social mining events in the world.

Share your expertise: if you want, you will be able to publish your products and services in the SoBigData Catalogue.

www.sobigdata.eu/associatemember

Next SoBigData events

ESME 2019 - Workshop

Ethical Social Mining and Explainability in AI

8th and 9th July 2019

Officine Garibaldi, Pisa, Italy

If you are interested in issues of ethics, the application of the GDPR and the right to explanation, in social mining and exploration context, and you want to participate in discussion with other experts of the topic, consider sending your application to ESME 2019 (Ethics, Social Mining and Explainability). The application form is open to both researchers and professionals. You can request the registration by filling out this form: goo.gl/RXWPXQ
The organizing committee will evaluate the applications; if you will be selected, the expenses for travel and accommodation will be covered by ProRes project (http://prores-project.eu/).

More info on page 29

Summer School on Analysing Disinformation

25th - 29th June 2019

Kings College London, UK

More info on page 30

International Data Science Summer School

2nd - 6th September 2019

Officine Garibaldi, Pisa, Italy

The Italian PhD programs in Data Science, offered in Pisa, Rome and Bologna, are jointly organizing the Data Science Summer School series, starting with the first edition to be held in Pisa from 2 until 6 September 2019. The summer school is aimed to reflect the interdisciplinary flavour of our PhD programs, offering lectures by high-level scholars from different domains, including data mining and big data analytics, machine learning and AI, network science and complex systems, digital ethics, computational social science and applied data science in general.

More info soon published on www.sobigdata.eu. Stay tuned!

Final SoBigData Conference

6th September 2019

Officine Garibaldi, Pisa, Italy

More info soon published on www.sobigdata.eu. Stay tuned!

Explainable Machine Learning: the new SoBigData exploratory

It is urgent to develop a set of techniques which allows the user to understand why an algorithm made a decision.

We are evolving, faster than expected, from a time when humans are coding algorithms and carry the responsibility of the resulting software quality and correctness, to a time when sophisticated algorithms automatically learn to solve a task by observing many examples of the expected input/output behavior. Most of the times the internal reasoning of these algorithms is obscure even to their developers. For this reason, the last decade has witnessed the rise of a black box society. Black box AI systems for automated decision making, often based on machine learning over big data, map a user's features into a class predicting the behavioral traits of individuals, such as credit risk, health status, etc., without exposing the reasons why. This is troublesome not only for lack of transparency but also for possible biases inherited by the algorithms from human prejudices and collection artifacts hidden in the training data, which may lead to unfair or wrong decisions. It is therefore urgent to develop a set of techniques which allows the user to understand why an algorithm made a decision.

http://www.sobigdata.eu/exploratories/explainable-machine-learning

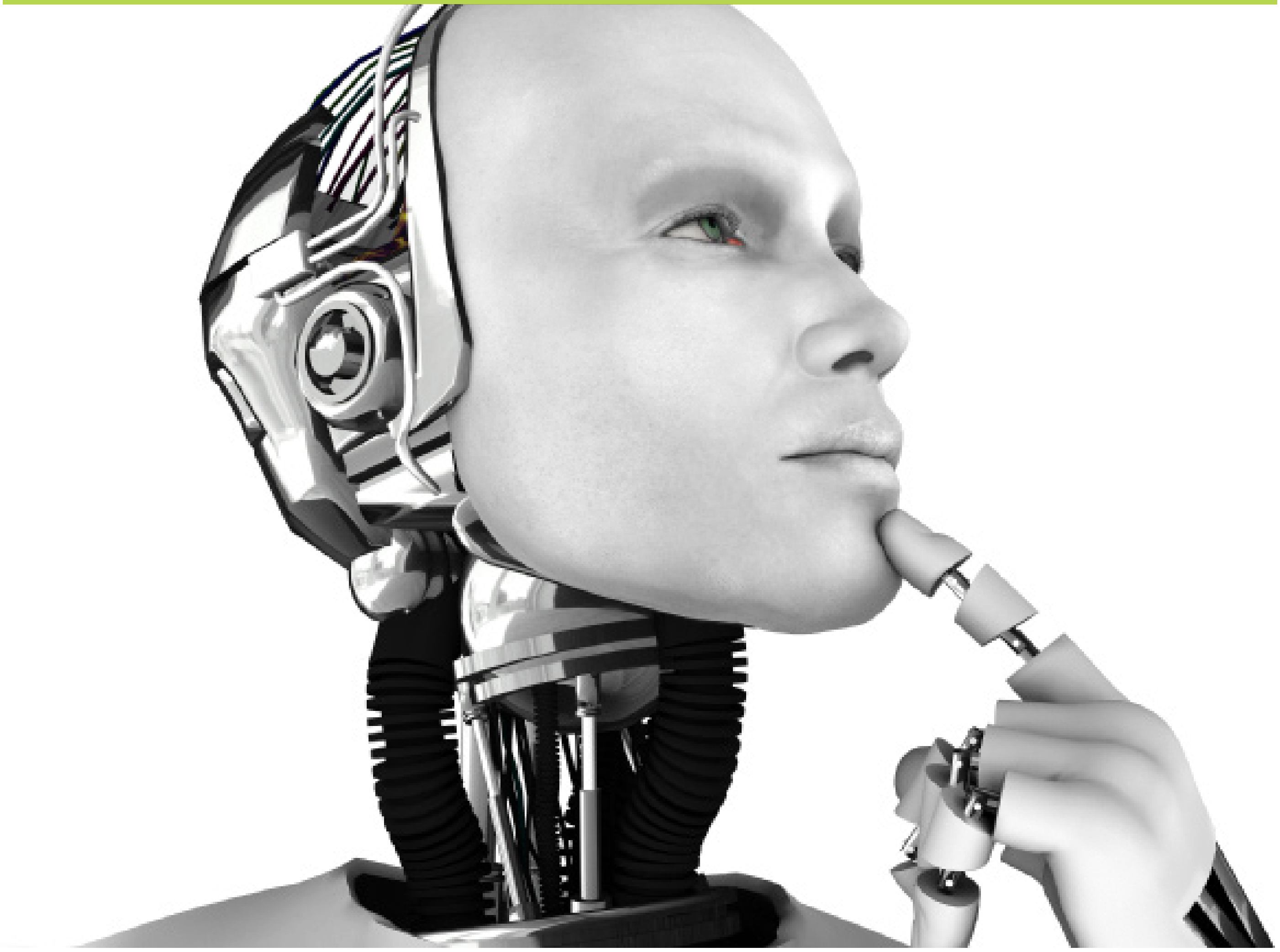


Image by Pngimg CC BY-NC 4.0

Transnational Access: an open call to visit the SoBigData R.I.

Interact with the local experts, discuss research questions, run experiments on non-public datasets and methods, present results at workshops/seminars.

The SoBigData project invites researchers and professionals to apply to participate in Short-Term Scientific Missions (STSMs) to carry forward their own big data projects. These opportunities are offered as part of SoBigData's Transnational Access (TNA) activities and calls for applications will be opened every six months.

We welcome applications from individuals with a scientific interest, professionals, startups and innovators that may benefit from training in data science and social media analytics. In order to apply you have to fill the Project Application Form.

Up to 4500 euros per participant, to cover all the cost incurred.

Funding for a short-term scientific mission (2 weeks to 2 months) is available up to 4500 euros per participant (to cover the cost of daily subsistence, accommodation, and economy flights). STSM bursaries are awarded on a competitive basis, according to the procedure described in the application pack and eligibility criteria below, and based upon the quality of the applicant, the scientific merit of the proposed project, and their personal statement.

PRE-REQUISITES

Good understanding of social data and, ideally, track record of prior social data analysis projects

Experience with using at least one of machine learning, natural language processing, and/or complex networks algorithms

THE GOAL

The goal is to provide researchers and professionals with access to big data computing platforms, big social data resources, and cutting-edge computational methods.

STSM visitors will be able to:
Interact with the local experts;
Discuss research questions;
Run experiments on non-public big social datasets and algorithms;
Present results at workshops/seminars.

The STSM visits will enable multi-disciplinary social mining experiments with the SoBigData Research Infrastructure assets: big data sets, analytical tools, services and skills.

RESEARCH CENTRES

Applications are invited for access at the following centres (infrastructures):

Gate *(Text and Social Media Mining), University of Sheffield;*

SoBigData.it *Pisa, Italy;*
Fraunhofer IGD, Darmstadt, Germany;

UT University of Tartu, Estonia;

L3S Research Center / Leibniz University Hannover;

Aalto University Finland;

ETHZ Zurich, Switzerland

The calls are published on the SobigData website and the main social channels.

For more info, please visit the SoBigData website (*www.sobigdata.eu*)

Open call
for SoBigData-funded TransNational Access
open call: continuous submission

The Fourth call on Transnational Access. Visit period: November 2018 / July 2019

Experiences from the TNA calls

A selection of research experiences based on the Transnational Access Calls.

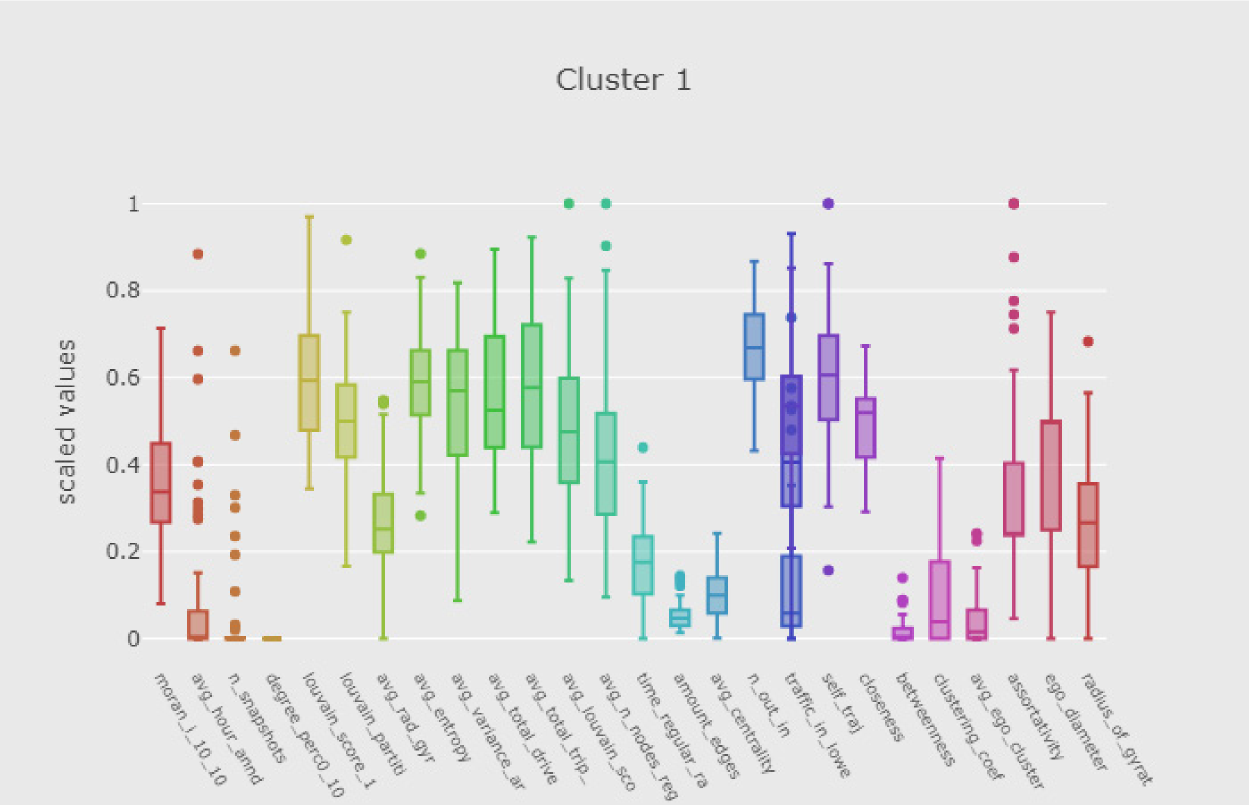
Semantics-enabled Transfer Learning for Mobility Analytics

Marta Sabou, Technical University of Vienna, TUWien | marta.sabou@ifs.tuwien.ac.at

Host: SoBigData.it | Pisa, Italy

A well-known drawback of supervised machine learning approaches is their dependence on training data, which is often very expensive to acquire. Transfer learning is an increasingly popular approach for overcoming the need for acquiring expensive training data in cases when supervised learning algorithms are applied across sufficiently similar domains or cases. In a nutshell, transfer learning methods allow the adaptation of models learned for one case (e.g., for classifying images of cats) to similar cases (e.g., for classifying images of dogs). A recent SoBigData TNA visit has focused on the topic of transfer learning of mobility analytics algorithms by making use of Semantic Web data. The visit took place at the KDDLab (http://kdd.isti.cnr.it/) of CNR Pisa and primarily involved work by Dr. Roberto Trasarti of the host institute and Dr. Marta Sabou (TU Wien), as visitor.

The problem: Mobility analytics algorithms have a wide range of applications in Smart Cities, which are investigated as advanced cyber-physical social systems (CPSS) as part of the CitySPIN project lead by Dr. Sabou. The focus of the visit was tailored to a particular aspect of mobility analytics, namely activity recognition in mobility data. Activity recognition refers to assigning a label to a movement based on its characteristics [1]. For example, given a raw movement trace (such as that collected from GPS traces), the goal is to segment this trace in locations and movement segments between those, and assign to each segment of this trace a suitable label that reflects the activity that the individual was undertaking during that movement segment. Such labels can, for example, reflect the purpose of motion, e.g., going home, going to work, shopping, leisure trips, picking up children etc. From the perspective of CPSS, such annotations of human activities are valuable for understanding the social component of CPSS and its behavior patterns, which, on its



turn will contribute to better adapting and coordinating the CPSS.

The National Research Council of Pisa has performed work on activity recognition in mobility data [1]. They learned a classification model (the ABC Classifier) which, given a raw trajectory, can classify segments of these trajectories in terms of the purpose of the journey (e.g., home, work, shopping). The classifier was trained on manually annotated GPS data from the city of Pisa and had a good performance, yet its application to raw GPS data from Florence lead to suboptimal performance. Obviously, Pisa and Florence are cities with diverse characteristics which lead to different mobility behavior patterns.

The question: In this context, the goal of the visit was to investigate the following questions: Is it possible to identify cities similar to Pisa on which the activity recognition algorithm would have a good performance? Could transfer learning perform better when applied to cities that are "similar" to the city on which the model was originally trained? How to use socio-demographic features of cities to enable transfer learning of mobility analytics algorithms?

The method: The visit focused on investigating these questions by taking an approach where semantic data from the Linked Open Data (LOD) cloud was collected about cities (e.g., geo-spatial features, statistical indicators about population etc) and used to cluster similar cities. Accordingly, the applied method encompassed the following steps:

1. City feature collection: Collect potential city characteristics/features from linked open data sources and also considering recommendations from literature on those features that most likely have an influence on individual mobility. City features were collected from the Linked Open Data site of the Italian National Statistics Institute (ISTAT, http://datiopen.istat.it/) and included: geo-spatial features (city surface, min and max altitude) and socio-demographic features related to the population of the cities. In particular, this dataset provides census data from 2011 for a variety of statistical indicators such as those related to their population (in terms of number, gender, age, occupation, education, number of foreign citizens).

2. City features preparation and selection: The features extracted are normalized and a correlation analysis is performed in order to select the maximum set of them which are not inter-correlated.

3. Clustering of cities: Cluster cities based on these characteristics. Particularly interesting is the cluster of cities similar to Pisa. Our hypothesis is that the adaptation of the ABC Classifier to these similar cities will be more successful.

4. City features versus mobility statistic: An analysis to detect how the mobility statistics extracted from GPS traces vary across the various clusters. From the original mobility dataset, for each city, we extracted statistical information (trip length, duration, speed) about the incoming, outgoing journeys as well as journeys inside the city. The goal was to analyze the differences between the distribution of values for all cities along these dimensions and that of certain clusters. This analysis step is currently ongoing.

5. Running and evaluating the ABC Classifier: Run the classifier on cities from the cluster of similar cities to Pisa as well as different cities from Pisa, and compare performance. The assumption is that performance will depend on how similar the cities are. Performance can in a first instance be estimated if the distribution of activity types is similar to that expected from the experiments on data from Pisa. Any large variations indicate a low performance.

The main output was the creation of a dataset of 522 Italian cities and their features extracted from LOD cloud (this data set is made available in the SoBigData catalogue). This dataset was also used to perform an initial clustering of cities, as per step 3 above. We relied on the support of the KNIME tool (see clustering workflow attached). We used k-Means Clustering (number of clusters 10, then 20) and manually inspected the results. We found that population size is one of the key differentiating features leading to clusters with very small (904) or large (271,767) populations. Within cities with similar population, the clustering often distinguishes between those with big/small surface. In the various clustering experiments Pisa and Florence are always in different clusters.

Results and Conclusions:
The major result of this visit consists in an approach to semantics-supported transfer learning for mobility analytics algorithms. We investigated this concept in a concrete case (that of the ABC classifier and with data about city features) leading to the following concrete results:
- A dataset of city characteristics extracted from linked open data;
- An Experiment for city-clustering.

The overall conclusion is that, semantic data (such as linked open data) is promising to support transfer learning in general, although this hypothesis still needs to be proved in future work which will be performed beyond the visit to investigate the performance of the activity recognition in cities classified as similar to Pisa. More broadly, this approach should be further investigated for other mobility algorithms as well as in other domains.

REFERENCES:
[1] S. Rinzivillo, L. Gabrielli, M. Nanni, L. Pappalardo, D. Pedreschi and F. Giannotti, "The purpose of motion: Learning activities from Individual Mobility Networks," 2014 International Conference on Data Science and Advanced Analytics (DSAA), Shanghai, 2014, pp. 312-318.

CitySPIN project: http://cityspin.net/

Future Perfect? The Social and Political Views of the New Tech Elite

Prof. John Torpey, City University New York, New York, USA

Host: SoBigData.it | Pisa, Italy

It is widely thought that innovations in information technology, nanotechnology, artificial intelligence, cognitive sciences, biotechnology, and other emerging technologies will fundamentally change society. But how? The goal of the project is to investigate how the members of the global high-tech elite envision the future of human society. The focus on tech elites derives from the assumption that their wealth, expertise, and control of technological developments gives them a pivotal role in contemporary social change. The high level questions faced by the project are about the social scenarios the members of the tech elite do foresee. What kinds of opportunities and threats do they anticipate? What values do they emphasize in public debate? What are their political views?

Problem and Research Questions

The problem posed by this project is to collect the information that is available on the Web with respect to statements (e.g., social posts, interviews...), actions (e.g., fundings, participation to events...) that better help to capture the position and the vision that the tech elite member have about social and political issues.
The idea is to collect such information and connect each person to each topic of interest for the project (described below) and to model their position on it. For example, what's the position of the tech elite on health? Is the tech elite focused on solving the current problems of a local regions, or is the tech elite aiming at producing breakthrough innovation for the whole planet? What the vision of the tech elite on education? Does the tech elite support more rewarding the best performing ones or promotes inclusion for all?

Context

The human language technologies group of ISTI-CNR in Pisa has a long standing research experience on text analytics problems, including text classification and information extraction. The TNA aims at supporting the investigation by providing automatic text processing method in order to gather and process the potential relevant information that is available on the Web.

Approach (Method and Data):

In order to explore these questions, we have identified CEOs and founders of the 100 largest tech enterprises (as measured by market capitalization) and we want to trace their conversations on such platforms as Twitter and Facebook, as well as on their personal, corporate, and foundation homepages for a period of approximately two months around the time of major political and tech-related events.

Preliminary activities:

A first non-automatic activity consisted in searching, for each of the 100 identified persons, their social profiles, any eventual foundation they control or any other source of information that directly reports on the thoughts and actions of the selected persons.
This manual data collection process was able to retrieve a Web presence of a foundation or similar organization that was connected to a tech elite person only for 24 cases. The search for social accounts resulted in even less result.
Other investigations on the Web led to the observation that the Web presence of a large majority of the selected person was so small to make the use of automatic processing method impractical. For example, a Google search of "Lin Xiucheng", ranked 54th on the list, only produces 6500 results, most of them related to short news and not reporting relevant information about the goal of

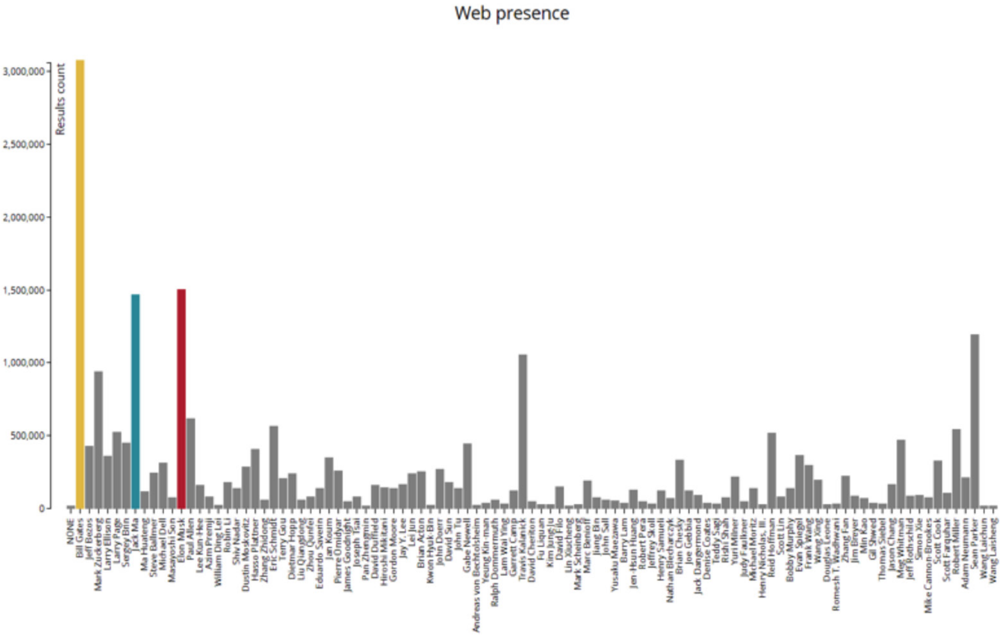


Figure 1

this project. This issue is not limited to person from a specific geographic area. For example, “Douglas Leone” a US venture capitalist, only appears in 21000 google search results. As reported in the graph below, the variance Web presence (using the number of results returned by a Google search as a proxy) of the selected persons is very high, with a large number of persons with minimal presence.

For many of these persons the information that is available on the Web is very generic (e.g., information about their wealth, their companies…) and only in minimal parts, when such information exists, reports on their direct action or thoughts about the topics of this project. These results made evident that the use of automatic text processing methods we originally envisioned for the project was not a viable solution.

For this reason we switched the main activity of ISTI-CNR into measuring how the Web represents and perceives the tech elite members. In this way, by enlarging the amount of data we can rely on, it has been possible to perform some statistical analysis on the collected data. Despite the obtained results cannot be claimed to model the personal views of the members of the tech elite, but only a perception of such views by a different population (the Web), they can be used as a potential validation source once the main goal of the project will be reached by using other methods (e.g., personal interviews).

In parallel with this activity, a second relevant activity has been the definition of relevant keywords that helped in selecting the main topics and keywords that describe which aspect of the social and political views we want to capture.

The result of this activity is the identification of seven thematic areas of interest:

- health
- education
- civil society
- social infrastructure
- social security
- environment
- art

Each area has been modeled with a set of relevant keywords. For example, for education some to the keywords are: Learning, Development, School, Teacher, Student, Degree, College, University, Research. For health some of the keywords are: Health, Cancer, Alzheimer, Disease, Cure, Medication, Vaccination, Epidemiology.

Also a number of dimension of analysis, with a binary characterisation, has been identified and modeled through keywords related to the two possible option of each dimension:

- goals: problem solving (e.g., solution, eradicate, fix) vs better world (e.g., create, invent, innovate).
- values: materialistic vs post-materialistic
- motivation: negative vs positive
- social dimension: inclusive, exclusive
- spatial dimension: local vs global
- temporal dimension: present vs future

A total of about 450 keywords have been produced.

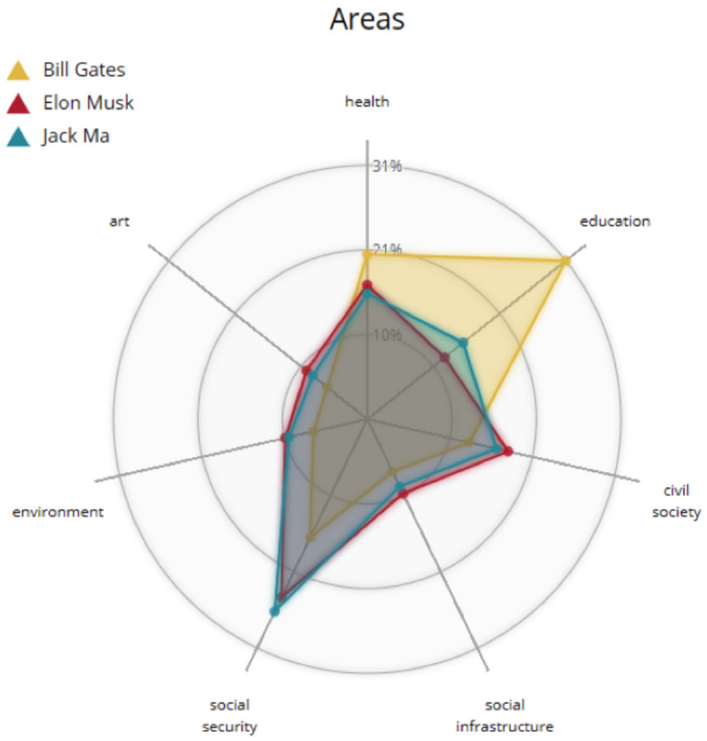


Figure 2

Data collection:

We issued more than 40k query to the Google search engine gathering the number of pages returned by each query.

Queries were formed by single entities, e.g., a person’s name “Bill Gates”, a keyword from a thematic area vaccination, or by pair of entities, e.g., “Bill Gates” vaccination.

Repeating this on all possible persons, keywords, and pairs - thus the large number of queries - it allowed us to measure the statistical correlation, by means of pointwise mutual information (normalized with a softmax), of persons with respect to the thematic areas and dimensions.

Visualization:

We collected the results of processing the search queries into an interactive visualization (hosted at https://www.esuli.it/demo/elite/).

A polar plot show the correlation between a selection (selectable by the user) of three persons with the thematic areas.

It is thus possible to directly compare the “profiles” of different persons.

The example in Fig.2 seem to indicate a higher correlation of Bill Gates with the themes of health and education than Jack Ma and Elon Musk, with the latter two sharing a similar profile.

A second plot then uses gauges to indicate on which side of every binary dimension a person is perceived to stay (the colors indicate the same person of the plot above).

Finally we applied a PCA process to the profiles determined by all the reported information in Fig. 3 (correlation to thematic areas and position in binary dimensions) to determine a similarity model among the tech elite persons. A dimensionality reduction allowed us to show such similarity model in a two dimensional plot (again the colors refer to the same persons of the above plots). The plot is interactive, hovering over a dot shows the name of the person. The plots are connected, clicking on a dot selects the person for the visualization the other plots.

Next Steps:

Even though the main goal of capturing the personal vision of tech elite members through their web presence proved to be a hard task, mainly due to the lack of public, reliable information for many of the identified subjects, this activity has been an opportunity to co-occurrence-based correlation methods to a novel tasks of profiling a subject against a set of topics of interest, and how to produce an effective visualization of the results.

Whenever the main goal of the project will be reached (other members of SoBigData have interest in providing support to this project), the comparison with the results obtained in this experience could provide additional insights on the differences between the tech elite vision and how this vision is perceived on the Web.

Dissemination:

The results from the statistical correlation study from web data has been published online (https://www.esuli.it/demo/elite/) and is available to be included into SoBigData exploratories.

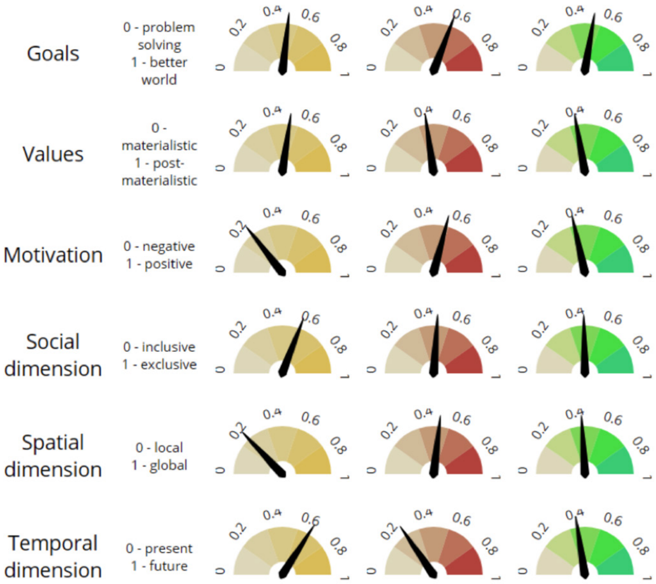


Figure 3

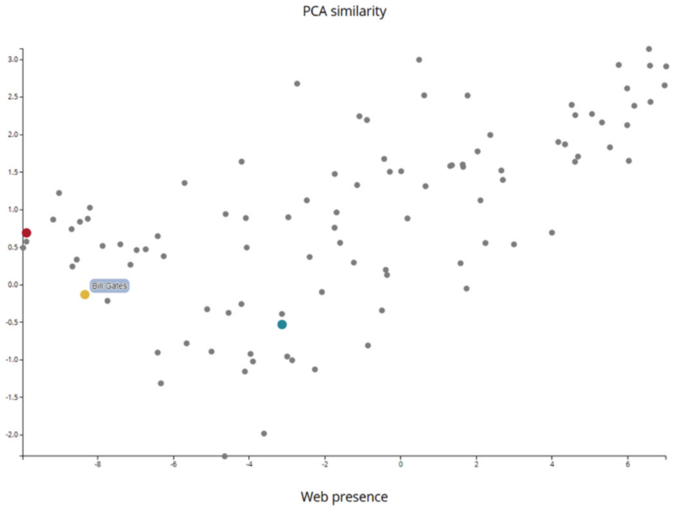


Figure 4

MAT-AD: A Conceptual Model for Multiple Aspect Trajectories with Analytical Dependencies

Prof. Ronaldo dos Santos Mello, Federal University of Santa Catarina (UFSC) - Brazil | r.mello@ufsc.br

Host: SoBigData.it | Pisa, Italy

Knowledge about people habits is of great value for smart city planners, public transportation administrators or companies. While for decades the approach to discover and measure the population profiles was through human surveys, nowadays, in the era of Big Data, we can infer much knowledge and the real behaviour about people from their everyday movement! The movement of an individual was first represented as a space-time track (a set of points) called Trajectory. Now, it became a multiple and heterogeneous n-dimensional track! See figure 1. It can show, for example, your movement: sometimes you`re walking, sometimes you are driving, sometimes you stop to work or to go to a restaurant, in some moments you`re posting in social networks, the weather is changing. This new kind of movement data is called Multiple Aspect Trajectory (MAT) as a lot of features (aspects) is related to your movement.

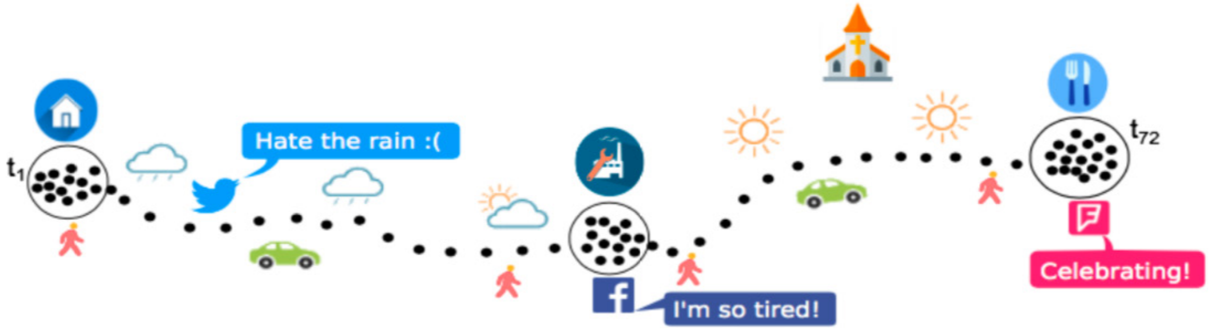


Figure 1

Trajectory data modelling and analysis are important research subjects because of the variety of information that may be extracted/inferred from these data, such as the daily habits of individuals obtained from their check-ins in social media. A conceptual model for representing data about a MAT was proposed by our research group. However, this conceptual model does not consider the representation of patterns discovered by data mining methods executed over MATs, which would be very useful for several applications that would like to have access to this hidden knowledge.

But how to model additional information about MATs for analytical purposes? One possible solution is to is to investigate MAT datasets to get insights about relevant analytical operations to be accomplished over them and define a conceptual modelling of MATs for analytical purposes from these operations and inferred requirements. That`s what we did by accessing the SoBigData City of Citizens exploratory: a large repository of people movement (https://sobigdata.d4science.org/group/cityofcitizens)! This kind of modelling for MATs is an open research issue and we call it MAT-AD (MAT-Analytical Dependency).

What is an "Analytical Dependency" (AD)? It means a dependency among instances of one or more aspects that is relevant for a given analysis. For example, it could be useful to verify if people that go to expensive restaurants (a place aspect) usually pay with a credit card (a way of payment aspect). Basically, an AD defines an implication X => Y, where X and Y can hold several constraints over aspects connected by logical operators. It looks like a classical association rule defined by data mining methods, but it is more simple than an AD in sense they define only conjunctive equality conditions over data attributes.

Some examples of relevant ADs we found during our analysis of City of Citizens data: (i) goal = 'Leisure (sport, excursion, ...) => mean of transportation = 'Motorcycle' OR mean of transportation = 'Automobile' (99% of confidence) (people prefer to have fun driving car or moto); (ii) goal = "Supermarket" => day_period = "6-12" OR day_period = "12-18" (99% of confidence) (people definitely do not go to the supermarket at night!).

These analyses were valuable in order to provide the foundations for the definition of an AD as well as MAT-AD: (i) an AD can hold a complex determinant and determined parts (as we said before), which generates a Constraint entity in our conceptual model; (ii) an AD is specific for a MAT domain entity (trajectory, moving object, point, etc), which generates the categories of ADs.

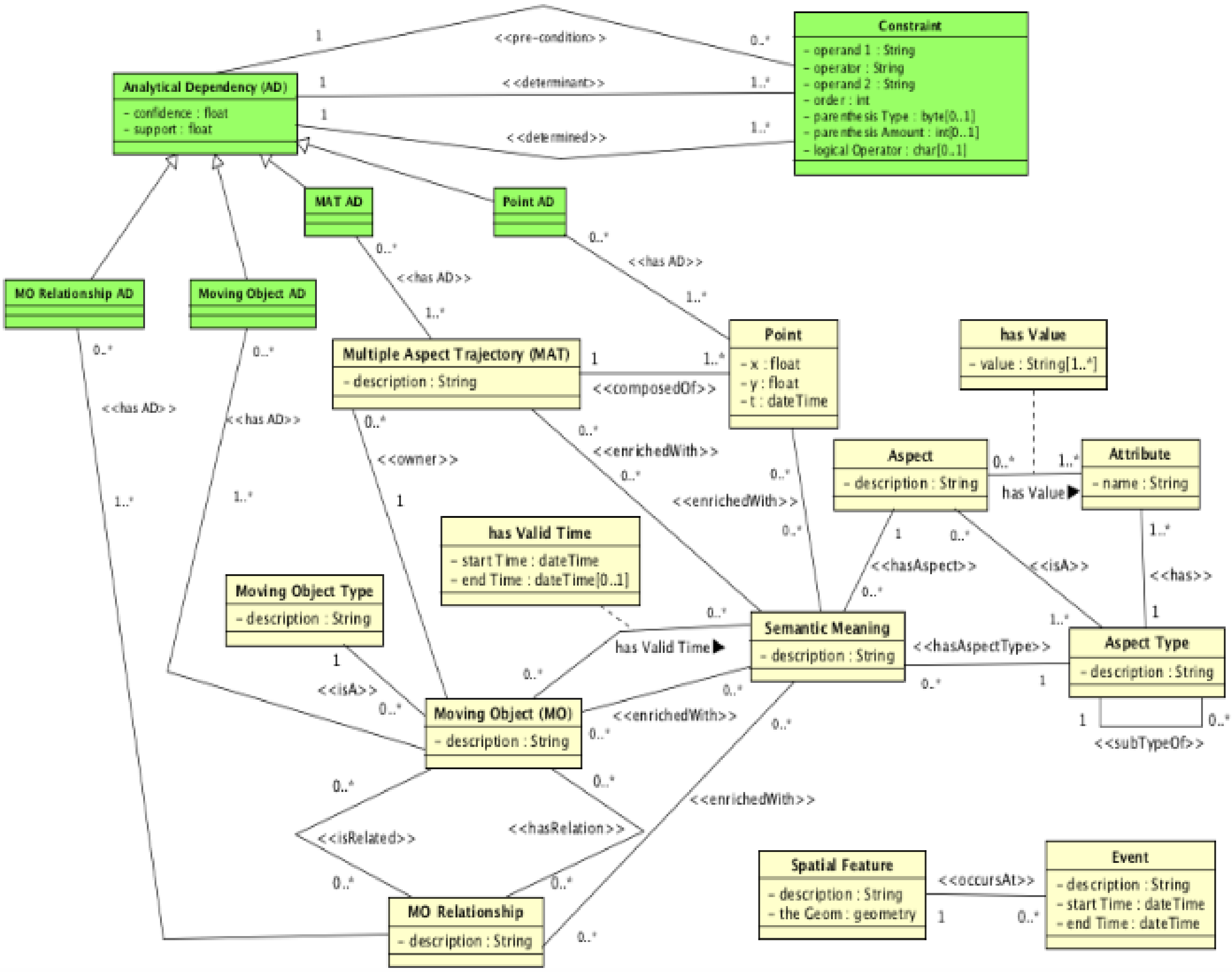


Figure 2 shows our MAT-AD conceptual model, the main contribution of our research. The yellow part represents the previous MAT conceptual model, and the green part represents the extension of MAT for representing ADs, with their categories and their constraints.

Using Machine Learning (Topic Modeling) to Define Product & Geographic Markets

Stephen Bruestle, Federal Maritime Commission | sdb8g@virginia.edu

Host: SoBigData.it | Pisa, Italy

Most economists do not use clustering. We like to identify a small, fixed number of parameters. And, clustering has many parameters.

Economics will have to change to stay relevant. Developments in data and data science make clustering more practical. Businesses are relying more on clustering. Economists should do the same.

My idea is simple. I will use clustering to solve an economics problem. Specifically, I will use clustering to define markets.

Defining markets is important to antitrust regulators. If you define a market too narrowly, then you get a monopoly. If you define a market too broadly, then you get perfect competition.

A good example of this is the recent Whole Foods merger case in the United States (U.S.). Whole Foods is the largest natural and organic grocery store chain in the U.S. They attempted to acquire Wild Oats, which is the second largest natural and organic grocery store chain in the U.S. The Federal Trade Commission tried to block the acquisition. They defined the relevant market as premium natural and organic grocery stores. Thus, the merger would create a monopoly. Whole Foods went to court to allow the acquisition. They claimed that their market includes the organic aisles of non-specialized grocery stores. Thus, the market would remain highly competitive (Carden, 2009).

Defining markets is important to businesses. For example, in India, the north and west are passionate about cricket. And the south is passionate about soccer. Therefore, the north and west see Adidas and Nike ads featuring cricket players. And the south sees Adidas and Nike ads featuring soccer players (Bhasin, 2017). Therefore, Adidas and Nike have found it useful to segment India into two markets: the north/west, and the south.

Thus, economists developed many statistical methods to define markets. Most of these methods are based on the following: If two goods are in the same market, then they are substitutes for one another. If the price of good A goes up, then consumers will demand more of good B. Therefore, the prices of goods A and B are correlated. And a shock to the price of good A will shock the consumption of good B.

Data scientists have developed a new kind of statistical method for automatically classifying things. They call this field: topic modeling. Usually, topic modeling algorithms classify documents based on the patterns of words in the documents. However, topic modeling has also been used in: classifying genetic sequences based on animal traits (Chen et al., 2010), object recognition in photographs (Fei-Fei and Perona, 2005; Sivic et al., 2005; Russell et al., 2006; Cao and Fei-Fei, 2007; Wang and Grimson, 2008), video analysis (Niebles et al., 2008; Wang et al., 2007), music analysis (Lawrence, 2009), and predicting user tastes and preferences (Marlin, 2004).

In my Short-Term Scientific Mission to Consiglio Nazionale delle Research (CNR) in Pisa Italy, I used topic modeling to classify consumer segments into markets based on their transactions.

In this project, I am developing a new application of Latent Dirichlet Allocation (LDA). LDA is the standard and the most common topic model. Blei et al. (2003) categorized documents based on the contextual patterns of text. I categorize consumer segments into markets based on the contextual patterns of purchases.

It makes sense to use LDA to define markets because LDA has three useful properties. First, I can motivate the LDA model as if it were an economics model for defining markets. The LDA model is general enough so that it applies to any market structure. Second, consumer segments can be classified in the same way that LDA classifies documents. It is an isomorphic problem, which means that the two problems have a structure-preserving one-to-one correspondence. Third, I can set up market data to fit LDA. Specifically, I am using data derived from SoBigData's Well Being and Economy Database.

The data is proprietary from Coop. Coop is one of the largest supermarket chains in Italy. The data is at the transaction-level. It includes prices and quantity of each product for each transaction. And it tracks consumers across multiple stores.

It is becoming more common for companies to gather this type of data. Yet, it is rare for a big corporation to allow academics to gain access to this type of data.

I also benefited from talking with the data scientists at CNR in Pisa.

Data scientists and economists have a lot in common. We both develop and apply statistics. We both answer policy relevant questions. We both use the scientific method and large datasets to understand how the world works.

The difference is the focus. Data scientists focus more on the empirics. Economists focus more on combining empirics and economic theory.

I found that my discussions with the data scientists at CNR in Pisa to be invaluable. They challenged me. This led to stronger economic theory.

I am eternally grateful. And, I hope to do more collaborations with data scientists in the future.

A Mixed Methods Approach to Crowdsourced Elections Data in Kenya

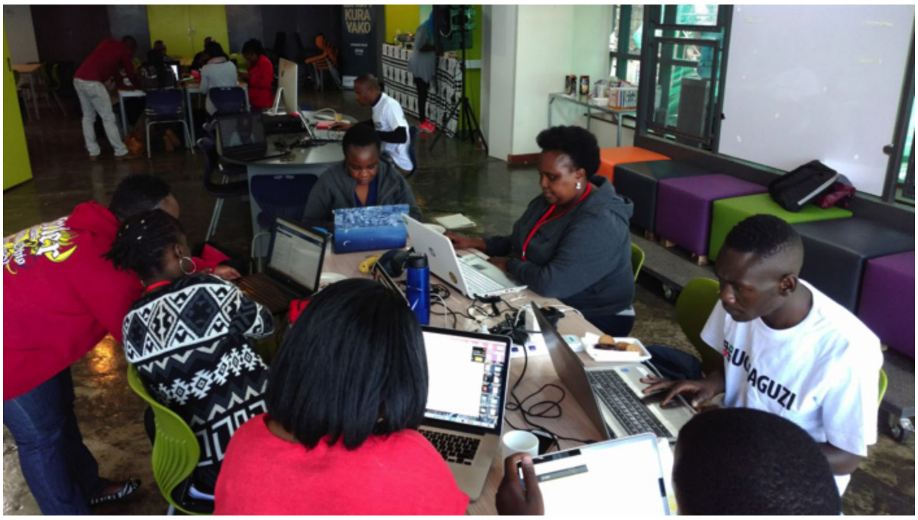
Shadrock Roberts, Ushahidi, Nairobi, Kenia | shadrock@ushahidi.com

Host: GATE | University of Sheffield, UK

The impetus for me to become a SoBigData fellow came to mind while I was monitoring the use of crowdsourcing for the Kenyan Presidential elections of 2017. The company I work for, Ushahidi, was born during the post-election violence of the 2007 Kenyan elections, when widespread violence and a government blackout to stifle information, were countered with the first version of the Ushahidi platform: used to collect testimony and eyewitness reports events unfolding throughout the country.

Ushahidi has implemented its software to help monitor every Kenyan general election since. As I was assisting with our monitoring project in 2017, I noticed that we had received only 4 reports for the neighbourhood of "Kibera," which is one of Africa's largest informal settlements and is often a flashpoint for tensions area during elections. I brief examination of Twitter revealed than 20 thousand mentions of the area. The discrepancy between what I was seeing on social media and what I was seeing in Ushahidi suggested that we were missing something big. The qualitative work I did in the field suggested that crowdsourcing of incoming reports might be reaching a limit that could be overcome with Natural Language Processing, which could provide an important way forward for Ushahidi. Our ability to serve the users of our software will benefit from improved methods in data science and social media analytics. My SoBigData project was to compare the Ushahidi and Twitter data sets to see which types of events each data set captured if they were comparable and if one could be used to say anything about the other.

Perhaps, unsurprisingly, the original Twitter dataset of 19,899 Tweets contained only 3,240 Unique Tweets mentioning "Kibera" created by 2,196 unique accounts. Of those accounts, only 21 (0.95% of total) Tweeted 10 times or more for a total of 412 Tweets (12% of all Tweets). While 74 accounts (3.4% of total) Tweeted 5 times or more for a total of 750 Tweets (22% of all Tweets). This is, more or less, a usual power-law distribution. The vast majority of these Tweets seemed to be little more than commentary (often partisan) and held a relatively small number of Tweets containing any relevant, on-the-ground, information.



The "Uchaguzi" monitoring team at work during the Kenyan General Election of 2017

The Ushahidi data, by contrast, was incredibly small: only 18 reports were published for the area under investigation. And of these, only 4 contained what might be considered "crisis events." From this, one could surmise that Ushahidi had, in fact, done an excellent job of filtering out much of the noise found in the Twitter data and presenting users with actionable information. However, when compared to qualitative data, it appears that the few "crisis events" that appeared in the Ushahidi data were already known to crisis responders. Nevertheless, I have yet to find any crisis events in the Twitter data that Ushahidi "missed" and it could be said that Ushahidi, at least, managed to present a view of what was happening in the ground which was free from so much of the noise found in the Twitter data.

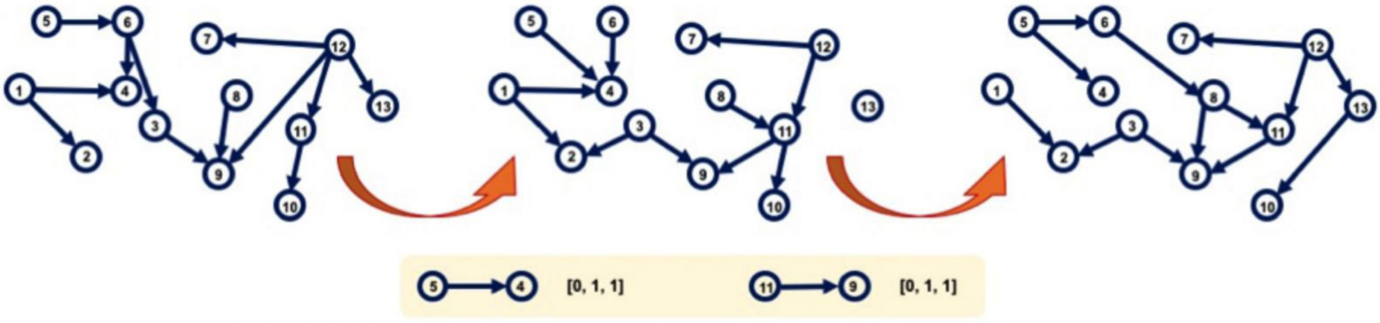
While I was not able to completely finish my analysis during my time with SoBigData, I was able to work with cutting edge techniques that allowed me to process data I wouldn't have been able to otherwise. I was also able to work with a team of experts who helped me extend my initial research protocol to continue mining the data for further insights. Thanks to the team at GATE and the University of Sheffield, I am currently trying to determine more about the unique accounts that were driving much of the Twitter conversation including what hashtags they were using and how these differed from Ushahidi; whether they were based in the area under investigation or foreign-based; and to what degree overt partisan sentiment can be detected in their commentary.

Unleashing the Power of Big Data to Build Smart Cities

Giulia Preti, University of Trento | gp@disi.unitn.eu

Host: Data Mining Group | Aalto University, Finland

The steady growth of the world population has increased the demand of resources like food, water, and energy, and thus providing a good quality of life and adequate services to the citizens has become a challenge for today's cities. The research community has given considerable attention to this problem, and has extensively used open data collected by public institutions and organizations for urban planning [1, 2], zoning regulation [3], public transport design, disease outbreak control, energy management [15], and disaster management [4, 5]. A large amount of this data is generated by sensor networks monitoring variables such as air quality, water quality, noise pollution, and road traffic load. These networks are dynamic by nature, as the value recorded by each node changes over time, new nodes can be added or removed from the network, and links can fail or be removed as well.



In this work, we focus on the detection of dense groups of correlated edges in dynamic networks, which are groups of edges that are topologically close and changed in a similar way in a period of time. These structures are important in root cause analysis and anomaly detection, among others. For example, a traffic accident affects the speed along neighbor roads, and thus it causes the appearance in the road network of a dense group of edges with similar levels of activity. Similarly, the disappearance of a dense group of edges in a mobile network can be caused by the failure of the base station serving the corresponding nodes.

Problem Definition

The input is a dynamic network, which is a series of labeled graphs where the set of nodes is fixed, while the edges can change both structurally (i.e., they can appear/ disappear) and qualitatively (i.e., they have a weight that changes over time). We propose two measures to compute the density of a group of dynamic edges. The first measure is the minimum average node degree in the subgraph induced by the edges measured in the snapshots of the network. The second measure, instead, is the average among the average node degrees in all the snapshots. Similarly, we propose two measure to express the temporal correlation of a group of edges. The first one is the minimum Pearson correlation between the series of values associated to the edges in the group, while the second one is the average Pearson correlation.

Then, given a density and a correlation threshold, our goal is to find all the maximal groups of edges with density and correlation greater than the respective threshold. We analyze four different formulations of this problem, obtained by pairing the four measures described above.

In addition, since some types of dynamic networks may naturally contain a large number of dense groups of correlated edges, we study also the problem of identifying a more compact subset of results that is representative of the whole set. In particular, given a threshold on the maximum pairwise Jaccard similarity between groups of edges, we want to find a set of groups with pairwise overlap less than the threshold.

Method

We developed an exact solution to two problem formulations, i.e., that based on the minimum correlation and the minimum density, and that based on the minimum correlation and the average density. The algorithm is a two-phase approach that first identifies maximal groups of correlated edges, and then extract those subgroups of edges that form a dense subgraph.

In the first phase, we create an auxiliary graph where nodes represent edges of the original graph, and links exist between edges having correlation above the given threshold.

We can easily show that a maximal clique in the auxiliary graph corresponds to a maximal group of correlated edges. In fact, since all the nodes in a clique are connected, the edges they represent have correlation above the threshold, and thus the minimum correlation value in the group is greater than the threshold. We used Min hashing [12] to efficiently generate the auxiliary graph, and a variant of the TAPER algorithm [2] to implement the search of maximal cliques.

In the second phase, we extract the connected components from all the cliques, retaining only those that are maximal to avoid redundant computations in the next steps. Then, we examine each component, starting from the largest ones to allow an early termination of the algorithm when a component is found to be dense. We recall that we are interested only in the maximal dense subgraphs.

If the density of a component is below the threshold, all its subcomponents are recursively examined. Flags are used to avoid the examination of the same subcomponent multiple times.

Results and Conclusions

We tested our algorithms on a real dynamic network created using Twitter data collected from 2011 to 2016 by filtering keywords related to the gun control, the abortion, and the Obamacare topic.

The preliminary results, obtained using samples of this network of increasing size, allowed us to identify the limitations of our exact solution and prove the effectiveness of the hashing technique used in the approximate solution. In particular, in the exact solution, the time required to generate the pairs of correlated edges increased exponentially with the size of the network, whereas a careful tuning of the parameters in the approximate solution led to accurate results in reasonable amounts of time.

The experiments conducted using different thresholds on the minimum correlation and minimum density, required for a group to be part of the result set, allowed us to compare the quality (i.e. interestingness) of the groups of edges found. In particular, they showed that a social network like Twitter tends to contain a lot of overlapping dense communities of correlated edges, which in some applications convey redundant information. This result therefore motivates the need for our problem formulation that has the goal of finding a compact subset of diverse groups.

Nonetheless, more experiments are required before assessing the importance of the results found and how they can be applied to real world problems.

Future Work

We plan to introduce additional optimizations to speed up our algorithms, and then perform further tests on a Mobile Phone network, as well as on synthetic graphs. The experiments on the synthetic graphs will help us to assess the effectiveness of our approaches. Finally, we plan to implement solutions to the remaining two problem formulations.

Acknowledgements

The research was conducted under the supervision of prof. Aristides Gionis at the DMG Group of Aalto University, Espoo, Finland, thanks to the SoBigData TNA program.

REFERENCES:
[1] Phithakkitnukoon, Santi, et al. "Activity-aware map: Identifying human daily activity pattern using mobile phone data." International Workshop on Human Behavior Understanding. Springer, Berlin, Heidelberg, 2010.
[2] Yuan, Yihong, and Martin Raubal. "Extracting dynamic urban mobility patterns from mobile phone data." International Conference on Geographic Information Science. Springer, Berlin, Heidelberg, 2012.
[3] Toole, Jameson L., et al. "Inferring land use from mobile phone activity." Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, 2012.
[4] Candia, Julián, et al. "Uncovering individual and collective human dynamics from mobile phone records." Journal of physics A: mathematical and theoretical 41.22 (2008): 224015.
[5] Traag, Vincent A., et al. "Social event detection in massive mobile phone data using probabilistic location inference." Privacy, security, risk and trust (PASSAT) and 2011 IEEE Third International conference on social computing (SocialCom). IEEE, 2011.
[6] A. Angel, N. Sarkas, N. Koudas, and D. Srivastava. Dense subgraph maintenance under streaming edge weight updates for real-time story identification. Proceedings of the VLDB Endowment, 5(6):574–585, 2012.
[7] P. Rozenshtein, N. Tatti, and A. Gionis. Finding dynamic dense subgraphs. ACM Transactions on Knowledge Discovery from Data, 11(3):27, 2017.
[8] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti. Event detection in activity networks. In Proceedings of the 20th ACM International Conference on Knowledge Discovery and Data Mining, pages 1176–1185, 2014.
[9] T. Fawcett and F. Provost. Activity monitoring: Noticing interesting changes in behavior. In Proceedings of the 5th ACM International Conference on Knowledge Discovery and Data Mining, pages 53–62, 1999.
[10] M. Mongiovi, P. Bogdanov, R. Ranca, E. E. Papalexakis, C. Faloutsos, and A. K. Singh. Netspot: Spotting significant anomalous regions on dynamic networks. In Proceedings of the 2013 SIAM International Conference on Data Mining, pages 28–36, 2013.
[11] M. lgorzata Steinder and A. S. Sethi. A survey of fault localization techniques in computer networks. Science of computer programming, 53(2):165– 194, 2004.
[12] J. Zhang and J. Feigenbaum. Finding highly correlated pairs efficiently with powerful pruning. In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pages 152–161, 2006.
[13] J. Chan, J. Bailey, and C. Leckie. Discovering correlated spatio-temporal changes in evolving graphs. Knowledge and Information Systems, 16(1):53– 96, 2008.
[14] Z. Guan, X. Yan, and L. M. Kaplan. Measuring two-event structural correlations on graphs. Proceedings of the VLDB Endowment, 5(11):1400–1411, 2012.
[15] P. Palensky, and D. Dietmar. Demand side management: Demand response, intelligent energy systems, and smart loads. IEEE Transactions on Industrial Informatics, 7(3): 381-388, 2011.

The SoBigData Soccer Data Challenge

Luca Pappalardo, KDD Lab, ISTI-CNR, Pisa, Italy | luca.pappalardo@isti.cnr.it

In soccer, a role is generally intended as a description of the behavior a player is expected to comply with during a match. A forward is expected to score goals or create goal opportunities, a defender is expected to prevent the opponents to score, a midfielder to act as a playmaker in the middle of the field. An interesting question that emerged recently in the sports analytics community is: How can we define a role in a fully data-driven way? Solving this problem is crucial to many actors in the soccer industry: soccer coaches, managers and scouts are continuously looking for data-driven tools to improve the retrieval of talented players with specific characteristics, for example to replace players or improve the quality of the club's roster.

The design of algorithms for the definition of data-driven roles has been the topic of the last Soccer Data Challenge (http://soccerchallenge.sobigdata.eu/), organized in Pisa by SoBigData on 12th and 13th october 2018. Selected among around 30 teams of "hackers" through a qualification phase, 11 finalist teams of programmers, scientists and soccer fans (50 people) competed morning and night during 30 hours to propose the most innovative solution to the aforementioned problem, exploiting the access to a unique dataset of match-event data which describe in detail all ball touches that occurred in the matches of the Serie A 2017/2018 (passes, shots, foulds, etc.).

In collaboration with Wyscout, the leading company on providing analytical services for soccer scouts, the Soccer Data Challenge provided some examples of real players playing in 19 wyroles, i.e., prototypical roles in soccer defined by soccer scouts of Wyscout. Then, the 11 finalist teams were requested to design an algorithm -- the wyrole detector -- to associate each player with a wyrole on the basis of the examples and the match event data provided. Once designed their wyrole detector, the 11 teams were requested to design a measure of player versatility, defined as a player's propensity to play in different wyroles in different matches or even during the same match.

Despite the difficulty of the proposed problem, the solutions provided by the teams, presented during a short talk in the presence of a jury of experts from both soccer industry and departments of computer science, were highly innovative and covered a wide range of possible solutions, from network theory to artificial intelligence. The team "Holly e Benjo", composed by PhD student at University of Siena, proposed a deep learning algorithm to associate a player with a wyrole based on their behavior on the field and was elected as the winner of the competition. An honorable mention goes to team "I Beppi", composed by underage people, which made a particularly impressive presentation.

The first edition of the Soccer Data Challenge was a big success with the public and the media, demonstrating the great and increasing interest around the emerging field of sports analytics. This motivated SoBigData to organize a second edition: even though the new problem that will be proposed is top secret now, the staff promises that professional soccer clubs will be present at the challenge, that there will be room for more teams, and that soccer players and coaches will involved. Looking forward to it.

Complexity 72h: 72 hours of Science

Angelo Facchini, IMT Lucca, Italy

From 7 to 11 May 2018 the first edition of the workshop Complexity72h took place in Lucca. Inspired by the 72h Hours of Science - whose participants were organized in interdisciplinary working groups aimed at carrying out a project in a three-day time, i.e. 72 hours - Complexity72h is an event aimed at bringing together young researchers coming from different fields of complex systems science: more than 30 young researchers joined the workshop, with the final goal of solving a data science problem and uploading a report of their work on arXiv by the end of the event. A team of tutors proposed a number of research themes and guided the groups through the development of the chosen projects. Alongside teamwork, participants attended lectures from scientists coming from different fields of complex systems and tutorials organized by the participants themselves, on topics like the description of the employed data sets, data visualization and complex network analysis with Gephi.

Lectures given by Fosca Giannotti (CNR Pisa, Italy), Claudio Tessone (University of Zurich, Switzerland), Chiara Poletto (INSERM, France) and Guido Caldarelli (IMT Lucca, Italy) provided a broad overview on topics of interest for data scientists, as the study of blockchain-based systems, the use of big data for analysing human dynamics, etc. Assigned projects covered a broad variety of topics as well: link prediction on social networks, geographic analysis of online-offline communities, the quantification of resilience of financial networks, the study of psychiatric disorders on brain networks.

The workshop was supported by the projects SoBigData and OpenMaker, granted by the H2020 programme.

At the end of the workshop the following papers were uploaded on the arXiv repository:

M. Cinelli, I. Echegoyen, M. Oliveira, S. Orellana, T. Gili, *Altered modularity and disproportional integration in functional networks are markers of abnormal brain organization in schizophrenia*, https://arxiv.org/abs/1805.04329 (2018)

S. Chakraborty, X. R. Hoffmann, M. G. Leguia, F. Nolet, E. Ortiz, O.Prunas, L. Zavojanni, E. Valdano, C. Poletto, *Dynamics of new strain emergence on a temporal network*, https://arxiv.org/abs/1805.04343 (2018)

M. Baltakiene, K. Baltakys, D. Cardamone, F. Parisi, T. Radicioni, M. Torricelli, J.A. van Lidth de Jeude, F. Saracco, *Maximum entropy approach to link prediction in bipartite networks*, https://arxiv.org/abs/1805.04307 (2018)

A. Bovet, C. Campajola, J.F. Lazo, F. Mottes, I. Pozzana, V. Restocchi, P. Saggese, N. Vallarano, T. Squartini, C. Tessone, *Network-based indicators of Bitcoin bubbles*, https://arxiv.org/abs/1805.04460 (2018)

D., Di Gangi, D. Lo Sardo, V. Macchiati, T. Minh, F. Pinotti, A. Ramadiah, M. Wilinski, G. Cimini, *Network sensitivity on systemic risk*, https://arxiv.org/abs/1805.04325 (2018)

A second edition of Complexity72h is going to be organized in Lucca in the days 17-21 June 2019.

Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter

Stefano Cresci, Wafi Group, IIT-CNR, Italy | stefano.cresci@iit.cnr.it

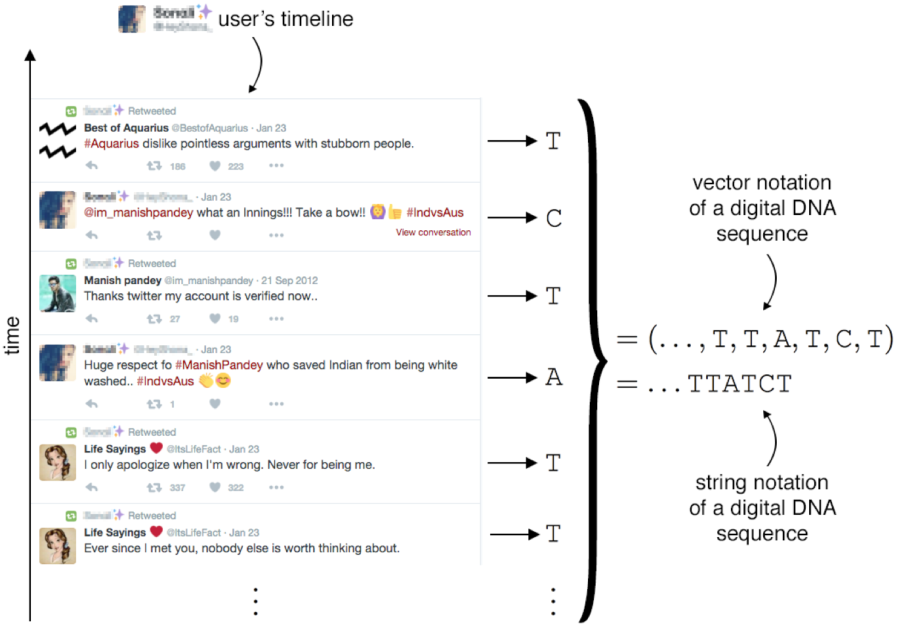
Beatrice Rapisarda, Wafi Group, IIT-CNR, Italy | beatrice.rapisarda@iit.cnr.it

Microblogs are increasingly exploited for predicting prices and traded volumes of stocks in financial markets. However, it has been demonstrated that much of the content shared in microblogging platforms is created and publicized by bots and spammers. Yet, the presence (or lack thereof) and the impact of fake stock microblogs has never systematically been investigated before. In the paper "Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter", we study 9M tweets related to stocks of the 5 main financial markets in the US. By comparing tweets with financial data from Google Finance, we highlight important characteristics of Twitter stock microblogs. More importantly, we uncover a malicious practice – referred to as cashtag piggybacking – perpetrated by coordinated groups of bots and likely aimed at promoting low-value stocks by exploiting the popularity of high-value ones. Among the findings of our study is that as much as 71% of the authors of suspicious financial tweets are classified as bots by a state-of-the-art spambot detection algorithm. Furthermore, 37% of them were suspended by Twitter a few months after our investigation.

> "Taking inspiration from biological DNA, we propose modelling online user behaviour with strings of characters representing the sequence of a user's online actions."
>
> S.Cresci

In order to understand this phenomenon, we deepen our previous analyses (in S.Cresci et al. *$FAKE: Evidence of spam and bot activity in stock microblogs on Twitter*) by performing a number of additional experiments on co-occurring cashtags, on financial markets, and on suspicious users:

- **we analyzed co-occurring cashtags** in financial tweets by focusing on their industrial and economic classification. In detail, we show that co-occurrences of stocks in suspicious tweets are not motivated by the fact that those stocks belong to the same industrial or economic sectors;
- **since real-world relatedness** (as expressed by industrial classification) is not a plausible explanation for co-occurring stocks, we then turned our attention to market capitalization. We demonstrate that, in suspicious tweets, high capitalization companies co-occur with low capitalization ones. Moreover, we show that this large difference cannot be explained by the intrinsic characteristics of our dataset, but it is rather the consequence of an external action;
- **we compared the social and financial importance** of investigated companies, highlighting that stocks of one specific market (OTCMKTS) feature


user's timeline
time
T
C
T
A
T
T
vector notation of a digital DNA sequence
$= (\ldots, \mathtt{T}, \mathtt{T}, \mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{T})$
$= \ldots \mathtt{TTATCT}$
string notation of a digital DNA sequence

[redacted] Retweeted
Best of Aquarius @BestofAquarius · Jan 23
#Aquarius dislike pointless arguments with stubborn people.
186 223

[redacted] · Jan 23
@im_manishpandey what an Innings!!! Take a bow!! #IndvsAus
View conversation

[redacted] Retweeted
Manish pandey @im_manishpandey · 21 Sep 2012
Thanks twitter my account is verified now..
27 19

[redacted] · Jan 23
Huge respect fo #ManishPandey who saved Indian from being white washed.. #IndvsAus
1

[redacted] Retweeted
Life Sayings @ItsLifeFact · Jan 23
I only apologize when I'm wrong. Never for being me.
337 322

[redacted] Retweeted
Life Sayings @ItsLifeFact · Jan 23
Ever since I met you, nobody else is worth thinking about.

Excerpt of a digital DNA extraction process in Twitter. In digital DNA each user action is associated to a given character, according to a predefined alphabet.

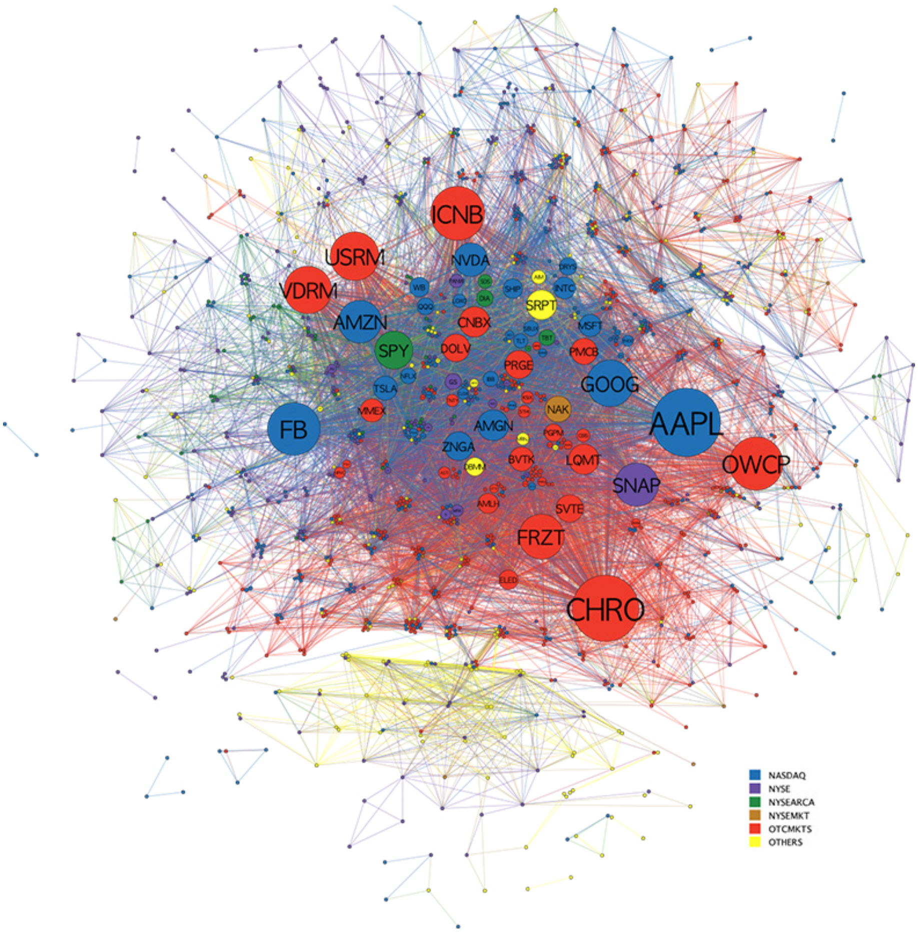
a suspiciously high social importance despite their low financial importance. This result is in contrast with measurements obtained for stocks of the other markets – e.g., NASDAQ , NYSE , NYSEARCA , and NYSEMKT;

- **we employed a state-of-the-art spambot detection technique**, called Digital DNA, to analyze authors of suspicious tweets. Results show that 71% of suspicious users are classified as bots. Furthermore, 37% of them also got suspended by Twitter a few months after our investigation.

Given the severe consequences that this new form of financial spam could have on unaware investors as well as on automatic trading systems, our results call for the prompt adoption of spam and bot detection techniques in all applications and systems that exploit user-generated content for predicting the stock market.

Analyses of suspicious users suggest that the advertising practice is carried out by large groups of coordinated social bots. Considering the already demonstrated relation between social and financial importance, a possible outcome expected by perpetrators of this advertising practice is the increase in financial importance of the low-cap stocks, by exploiting the popularity of high-cap ones.

The potential negative consequences of this new form of financial spam are manifold. On the one hand, unaware investors (e.g., noise traders) could be lured into believing that the social importance of promoted stocks has a basis in reality. On the other hand, also the multitude of automatic trading systems that feed on social information could be tricked into buying low-value stocks. Market collapses such as the Flash Crash*, or disastrous investments such as that of Cynk Technology**, could occur again in the future, with dire consequences. For this reason, a favourable research avenue for the future involves quantifying the impact of social bots and microblog financial spam in stock prices fluctuations, similarly to what has already been done at the dawn of financial e-mail spam.



Cashtags Co-Occurrences

*https://www.telegraph.co.uk/finance/financial-crime/11553696/What-happened-during-the-Flash-Crash.html

**https://www.huffingtonpost.com/2014/07/10/cynk-technology-stock_n_5573862.html

REFERENCES:

Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2018). Cashtag piggybacking: uncovering spam and bot activity in stock microblogs on Twitter. CoRR, abs/1804.04406.

Stefano Cresci, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi. 2018. $FAKE: Evidence of spam and bot activity in stock microblogs on Twitter. In Proceedings of the 12th International Conference on Web and Social Media (ICWSM'18). AAAI.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2016. DNA-inspired online behavioral modeling and its application to spambot detection. IEEE Intelligent Systems 31, 5 (2016), 58–64.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm shift of social spambots: Evidence, theories, and tools for the arms race. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW'17 Companion). ACM, 963–972.

Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2018. Social Fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. IEEE Transactions on Dependable and Secure Computing 15, 4 (2018), 561–576.

This work is developed in the context of Societal Debates exploratory in SoBigData. More information can be found in www.sobigdata.eu

Refugee migration: how the refugee crisis is perceived in Europe

Cristina Muntean, HPC Lab, ISTI-CNR, Italy | cristina.muntean@isti.cnr.it

We are recently witnessing one of the largest movement of migrants and refugees from Asian, African and Middle- East countries towards Europe. The United Nations High Commissioner for Refugees (UNHCR) estimates one million of refugees arrived at the Mediterranean coasts in 2015 mainly from Syria (49%), Afghanistan (21%) and Iraq (8%). The largest wave of arrivals started in August 2015 following a main route through Turkey, Greece, Macedonia, Hungary and Austria to Germany, France, UK and other northern European countries.

The implications of this refugee crisis are complex. The whole phenomenon is nowadays object of a heated and polarized debate. Understanding how the debate is framed between governmental organizations, media and citizens may help to better handle this emergency. To this end, we made a study on Twitter about the perception of the refugee crisis in Europe, published at ASONAM 2016 *[1]*.

Through the analysis of the Twitter online social network, we address the following questions: “How is the European population perceiving this phenomenon? What is the general opinion of each country? How is perception influenced by events? What is the impact on public opinion of news related to refugees? How does perception evolve in time in different European countries?”

We propose an analytical framework able to investigate discussions about polarized topics in online social networks from many different angles. The framework supports the analysis of social networks along several dimensions: time, space and sentiment. The steps are the following: 1) extract relevant spatial information, 2) enrich data with the sentiment of the message and of the user (retrieved in an automatic iterative way), 3) perform multidimensional analyses considering content and locations in time. The approach is general and can be easily adapted to any topic of interest involving multiple dimensions. It is scalable due to the automatic enrichment procedure. For the scope of this paper, we use our framework to outline the European perception of the refugee crisis.

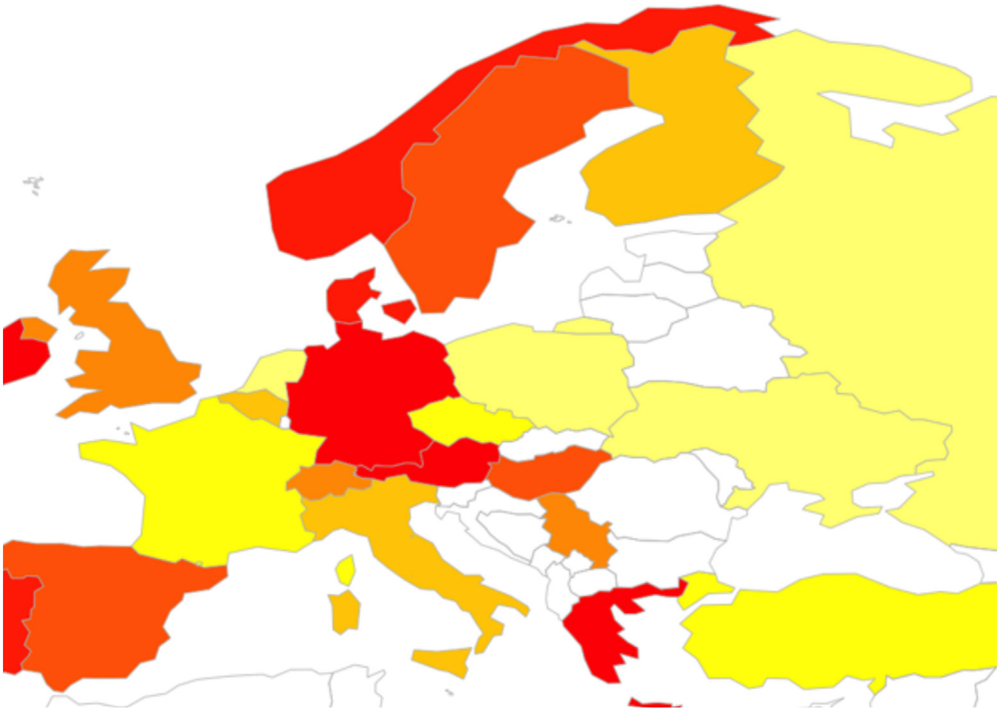


Fig. 1: Index ρ across European countries: red corresponds to a higher predominance of positive sentiment, yellow indicates lower ρ.

We used the Twitter Streaming API to collect English tweets data under the Gardenhose agreement (10% of all tweets on Twitter) in the period from mid-August to mid-September 2015 out of which we selected the tweets related to the refugee crisis topic, called the relevant tweets.

For each tweet, we extract two kinds of spatial information if present: the user location of the person posting the message and the mentioned locations within the tweet text. The user location is structured in two levels, the city (if present) and the country. The user city is identified from the GPS coordinates or place field when available. Since GPS and place data are quite rare we used the free-text user location field to enrich location metadata. We identified locations in the user-generated field based on location data from the Geonames dictionary which fed a parsing and matching heuristic procedure. The user country is collected in a similar way and when not explicitly present we infer from the city. The mentioned locations in the text are also represented at city and country level, and they are extracted from tweets’ text with the same heuristic procedure as for user location.

We are interested in understanding if the user has a positive feeling in welcoming the migrants or if he/she mainly expresses negative feelings (fear, worry, hate). Therefore, the dataset is enriched with information about the sentiment for both of tweets and users.

We consider two polarized classes Pro refugees (c+) and Against refugees (c−). We implemented the algorithm PTR (Polarization Tracker) *[2]* to assign a class to each polarized tweet and to each polarized user in an iterative way by considering his/her tweets and the hashtags contained. The approach proposed is suitable to track polarized users according to a specific topic which is in our case the "refugees phenomenon".

The procedure adds information about polarization of the users by polarized hashtags extension through the analysis of all the tweets written by an already polarized user and not only the polarized tweets. The iterative procedure is run until convergence has been reached. The combination of location and sentiment is done by crossing the space and sentiment.

From the analysis of the extracted hashtags, we can see that people with a positive sentiment prefer to use the term refugees, while people with a negative sentiment refer to them as migrants, thus minimizing the fact that they are escaping war and persecution. Users with a negative sentiment frequently use refugees and the Islamic religion together, somehow correlating, in a prejudicial way, refugees with Islam and terrorism. We observe that individuals with negative sentiment are often patriotic and not pro Europe.

Our study is driven by the analytical questions below:

-What is the evolution of the discussions about refugees migration in Twitter?
-What is the sentiment of users across Europe in relation to the refugee crisis? What is the evolution of the perception in countries affected by the phenomenon?
-Are users more polarized in countries most impacted by the migration flow?

In the paper, we show that the proposed analytical framework and the methodology can be used to mine knowledge about the perception of complex social phenomena. Our study shows differences in positive and negative sentiment in EU countries (Fig. 1), in particular in the UK (Fig. 2), and by matching events, locations and perception, it underlines opinion dynamics and common prejudices regarding the refugees.
The analysis revealed that European users are sensitive to major events and mostly express positive sentiments for the refugees, but in some cases, this attitude suddenly changes when countries are exposed more closely to the migration flow.

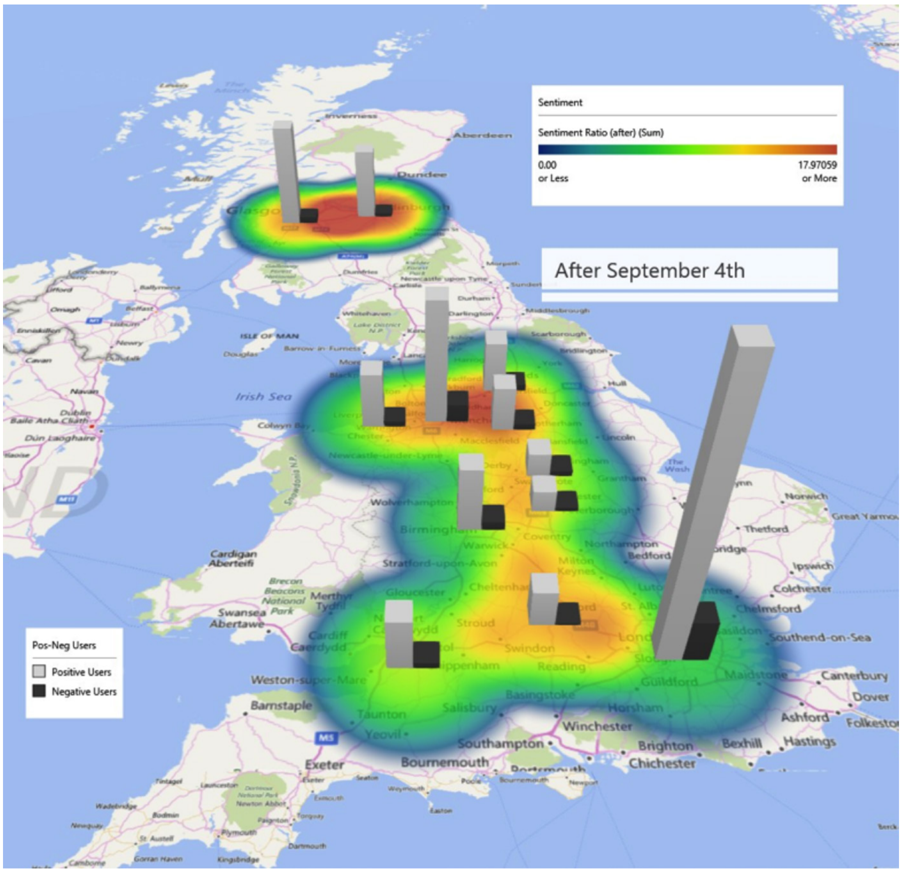


Fig 2. Positive and negative users for different cities in the UK in all period after September 4. In the infographic, the pies/bars show the number of polarized positive and negative users by city and the heat map in the background indicates the value of ρ for the cities considered in the legend.

As future work we intend to adapt the framework to a real-time streaming scenario and to add more dimensions such as the type of user and the network relationships in the Twitter user graph.

REFERENCES:
[1] Mauro Coletto, Andrea Esuli, Claudio Lucchese, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, Chiara Renso: "Sentiment-enhanced multidimensional analysis of online social networks: Perception of the mediterranean refugees crisis." ASONAM 2016: 1270-1277

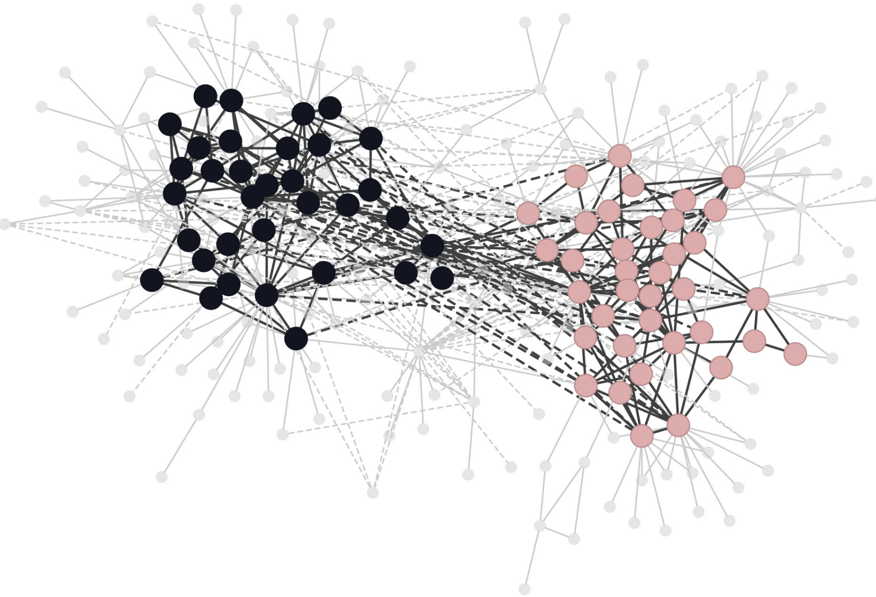
[2] M. Coletto, C. Lucchese, S. Orlando, and R. Perego, "Polarized user and topic tracking in Twitter," in SIGIR 2016, Pisa, Italy, 2016.

This work is developed in the context of Migration Studies exploratory in SoBigData. More information can be found in www.sobigdata.eu

The detection of polarization in social networks

Bruno Ordozgoiti, Aalto University, Finland | bruno.ordozgoiti@aalto.fi

The rapid expansion of social media platforms in recent years has led to a dramatic increase in the scale and pace of public debate. In this context, polarization on controversial issues is becoming a major concern. Even though a healthy amount of controversy can stimulate fruitful discussions, excessive contention can be harmful to individuals and society as a whole. Over the last decade, we have witnessed how polarization in social media is frequently associated with online abuse and the adoption of extreme ideologies. This, together with the problems posed by the appearance of bots, targeted advertisement campaigns and the spread of misinformation, has become one of the key challenges to be solved in order to ensure the healthy evolution of these platforms and their role in our society. This has led to the emergence of numerous research initiatives in this direction.



An example of two hidden polarized communities in a network representing interactions between US congresspeople. Solid edges are positive, while dashed edges are negative. The vertices in grey do not participate in any of the two polarized communities: either they have too few connections with the communities, or the polarity of their relations is mixed, so their position within the debate is unclear.

Polarization has received a great deal of attention in political and social sciences in the past. However, social media platforms have taken this phenomenon to an entirely different scale, providing an unprecedented source of valuable data. As well as creating opportunities for innovative analyses and insights, this circumstance brings about novel algorithmic challenges.

One of the fundamental data-analysis tasks in this context is the detection of polarization in social networks. The Data Mining group at Aalto University has recently been focusing on this issue, and in particular, the problem of discovering polarized communities in signed networks.

Social networks can be represented as graphs, that is, sets of nodes connected by edges. For example, we can encode the Facebook friendship network by representing each user as a node, adding edges between users who are friends with each other. Signed networks extend this representation by simply annotating each edge with either a positive or a negative sign. In a social network, we can interpret these signs as signifying either friendly or antagonistic interactions between users.

The bulk of the literature on finding antagonistic communities in signed networks focuses on partitioning the graph, that is, assigning each node to either of two -or more- groups. However, in the context of social networks, two polarized communities can be hidden within a sizable body of other nodes, which might be neutral to the debate taking place or indifferent to the ideological positions that divide the opposing groups. Therefore, in contrast to previous approaches, we propose a method to locate two polarized communities of relatively small size, leaving out nodes that do not side clearly with either faction. In addition, our algorithms can be tuned to control the size of the located groups, constituting a valuable addition to the social-media analyst's toolbox.

Our approach relies on spectral methods, which means that we can exploit existing, highly optimized, software packages for matrix computations. This, in combination with the sparse connectivity patterns that usually characterize social networks, allows us to efficiently locate polarized communities in networks comprised of millions of nodes.

An extensive set of experiments on a variety of real networks showed that polarized structures can indeed be located using the proposed techniques.

This work is developed in the context of Societal Debates exploratory in SoBigData. More information can be found in www.sobigdata.eu

Temporal mixture models: a case study on Bitcoin market

Tian Guo, ETH Zurich, Switzerland | tian.guo@gess.ethz.ch

Antulov-Fantulin Nino, ETH Zurich, Switzerland | nino.antulov@gess.ethz.ch

For a prediction task with data from different sources, these multi-source data interact to drive the fluctuation of the target variable. Modeling and capturing the interaction between these data is not only important for prediction, but also for interpreting the contribution of different sources to the prediction.

In this work, we study the proposed temporal mixture models by using the data from a rapidly rising area, Bitcoin [1, 2]. It is well known that Bitcoin is the first decentralized digital cryptocurrency, which has shown significant market capitalization growth in the past few years. It is important to understand what drives the fluctuations of the Bitcoin exchange price and to what extent they are predictable.

We focus on the ability to make the short-term prediction of the exchange price fluctuations (measured with volatility) towards the United States dollar by using the data of buy and sell orders collected from one of the largest Bitcoin digital trading offices in 2016 and 2017. We construct the temporal mixture model over the volatility and trade order book data, which is able to outperform the current state-of-the-art machine learning and time series statistical models.

With the gate function of our temporal mixture model, we are able to detect regimes when the features of buy and sell orders significantly affect the future high volatility periods. Figure 4 illustrates our model with a toy example. Figure 5 shows the examples of dynamical importance of the order book features and volatility history over the prediction.

This work demonstrates the prospect of the temporal mixture model as an accurate and interpretable forecasting model over multi-source data.

REFERENCES:
[1] Guo T, Bifet A, Antulov-Fantulin N. Bitcoin volatility forecasting with a glimpse into buy and sell orders[C]//2018 IEEE International Conference on Data Mining (ICDM). IEEE, 2018: 989-994.

[2] Guo Tian, Antulov-Fantulin Nino. "Predicting short-term Bitcoin price fluctuations from buy and sell orders." preprint arXiv:1802.04065, 2018.

This work is developed in the SoBigData context. More information can be found in www.sobigdata.eu

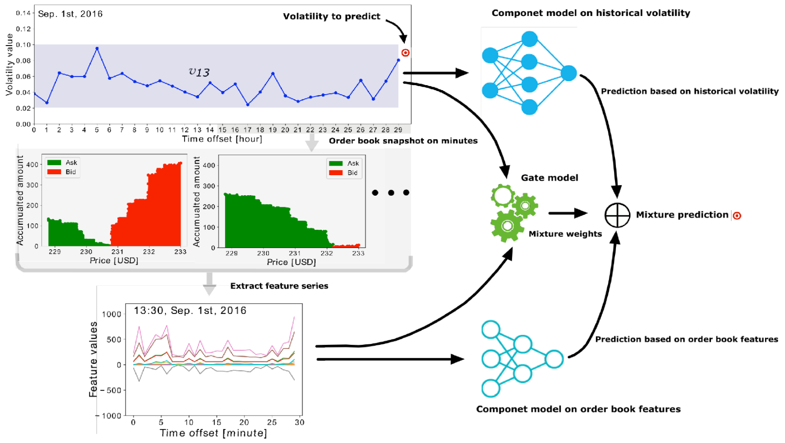


Figure 1: Temporal mixture model for volatility prediction. A top example: in order to predict the volatility denoted by the red target in the top panel, two component models respectively consume historical volatility (i.e. in the blue area) and order book features extracted from order book snapshots. Gate model utilizes both historical volatility and order book features to learn mixture weights on component models. The prediction is obtained by weighted sum of component predictions.

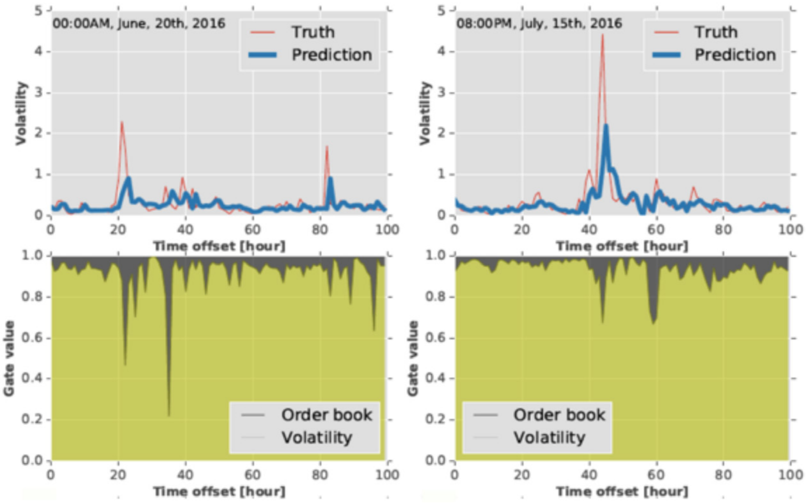


Figure 2: Evaluation results. Top panel: true values and predictions of two sample testing periods. Bottom panel: mixture weights of components respectively on historical volatility and order book features. It is observed that order book contributes more to high volatility with large gate values.

Training Material for Business Data Analytics

Marlon Dumas , Institute of Computer Science, University of Tartu, Estonia | marlon.dumas@ut.ee

Rajesh Sharma, Institute of Computer Science, University of Tartu, Estonia | rajesh.sharma@ut.ee

The demand of Data Scientists has increased in the past years, however, the supply to this demand is very low. This is partially due to the shortage of teaching experts and training material. As the subject is relatively new, thus, not all universities are able to attract data scientists who can teach the subject.
To fill this gap, the university of Tartu has created training material on Business Data Analytics. The material is publicly available on the web. That is anyone who has access to the internet can access it. This course is meant for students as a hands-on experience for solving business problems in the fields of sales, marketing, and business operations by applying statistical analysis and data mining techniques.

This course can be categorized as more of problem-oriented rather than solution-oriented. In every new lecture a new problem is introduced from the domain of business analytics keeping customers as a main central entity. Each lecture is also complemented with the use cases and solutions from data science field. Each lecture consists of 4 hours, where in the first 2 hours, the theoretical concepts are discussed, followed by 2 hours of lab session, where implementation details of various algorithms are discussed using the R programming language.

The course can be divided into three parts. **The first part** provides a basic introduction and background with respect to business data analytics, including the motivation and the importance of business data analytics by discussing various use cases. Next, an overview of various data descriptive and visualization methods is discussed.
The second part of the course is about various machine learning solutions with respect to customer related problems. The problems include customer segmentation and, customer life cycle management and value. In the later part, we discuss solutions about regression and churn problems. In addition, we also discuss the problems of up-selling and cross-selling for boosting the sales.
The third part of the course includes various other problems which exploits non machine learning techniques such as social network analysis, sentiment analysis, time series analysis for problems related to brand value monitoring, product diffusion, A/B testing, financial forecasting to name a few.

Total visits of the online material hosted at the University of Tartu is 67069 (based on Google analytics) for the last 18 months. A version of this course has been delivered at International School of Economics Tbilisi (ISET), Georgia, which is well received by students there as well. University of Tartu is particularly thankful to H2020 SoBigData grant in providing financial support for developing this course.

This work is developed in the SoBigData context. More information can be found in www.sobigdata.eu.

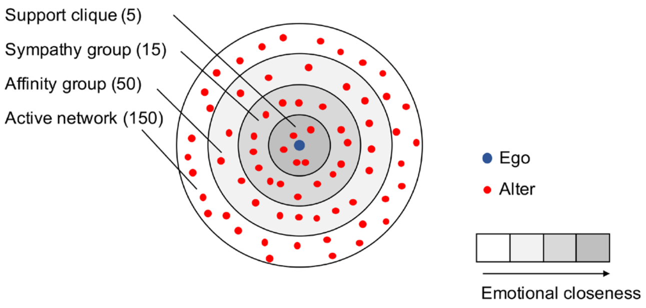
The limits and impact of friendship in online social networks

Andrea Passarella, Ubiquitous Internet Group, IIT-CNR, Italy | a.passarella@iit.cnr.it

Chiara Boldrini, Ubiquitous Internet Group, IIT-CNR, Italy | chiara.boldrini@iit.cnr.it

Personal networks, i.e., the ensemble of social relationships that an individual entertains with other people, have a significant influence on the quality of life of the individual in terms of, e.g., job opportunities, social support, power and influence in organizations. Personal networks are closely related to the concept of social capital, i.e., the network of connections, loyalties, and mutual obligations that translates into favors and preferential treatment.

Personal networks are studied using the graph abstraction of ego networks, which focus on the social relations between an individual (ego) and its peers (alters). An edge represent a relationship and it's typically weighted. The ego networks of our offline relationships (i.e., considering in real life interactions) are structured around nested groups: the group size increases but the emotional closeness decreases as we move outwards in the ego network (Figure 1). Thus, our spouse and best friend are expected to be found in the innermost layer, while our casual acquaintances are relegated in the outermost one. This layered structure emerges because humans have a finite brain capacity to allocate to nurturing meaningful social relationships. In addition, the number of social relationships that an individual can actively maintain is finite and around 150 alters, on average. This number is known as Dunbar's Number.



Ego network layered structure

Anthropologists and sociologists have known for a while that many traits of the ego's offline social behavior are determined by the ego network structural properties: for example, how people disseminate information, how they collaborate with each other, or how willing they are to share resources with others. But what happens when online social relationships are considered instead of in real life ones?

We have carried out a set of studies using data from both Facebook and Twitter and we have found that many properties carry over from the offline to the online world: social relationships are still organized around a layered structure and the size of each layer is similar. However, using the online data we uncovered the presence of an additional layer, in the innermost part of the personal network, which comprises at most only one or two people. This layer has been postulated since long in the anthropology literature, but data had never been fine grained enough to reveal its existence. Using online social networks data, we have possibly provided for the first time ever empirical evidence about this sociological hypothesis, confirming that online social networks are an extremely valuable microscope for studying human social relationships.

Another long-standing popular notion in the offline social network domain is that of the six degrees of separation, whereby a piece of information can be delivered to any selected recipient using at most five intermediaries. Again, how will this finding carry over from the offline to the online world? Information diffusion has been studied on the Facebook graph and researchers have found that any two Facebook users can be connected with 4-5 degrees, at most. These results have been obtained assuming that all social links are equivalent, i.e. expecting that people will be as willingly to be intermediaries for close friend as much as for casual acquaintances. However, not in all applications this might be the case (consider, as an example, information diffusion in which the piece of information occupies a non-negligible amount of space on the intermediaries' smartphones). In this scenario, people that have stronger social relationships are expected to help each other at the expense of their weaker social links. We have studied the impact of this trusted information diffusion relying on data from Facebook and we have found that when people diffuse information based on the strength of their social relationships, the degrees of separation are much more than 6 (and as high as 17 in some extreme cases.)

REFERENCES:

Dunbar, R. I. M., Arnaboldi, V., Conti, M., & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. Social Networks, 43, 39–47. https://doi.org/10.1016/j.socnet.2015.04.005

Arnaboldi, V., Conti, M., Passarella, A., & Dunbar, R. I. M. (2017). Online Social Networks and information diffusion: The role of ego networks. Online Social Networks and Media, 1, 44–55. https://doi.org/10.1016/J.OSNEM.2017.04.001

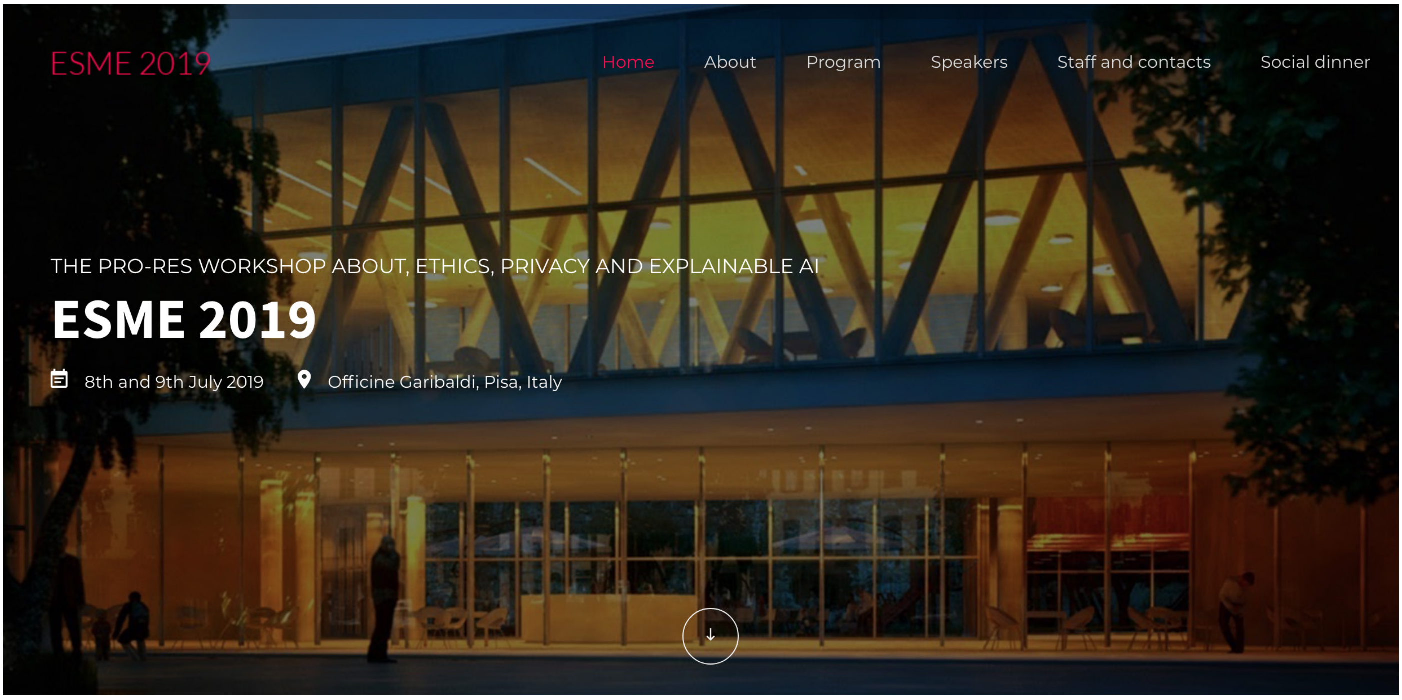
This work is developed in the context of Societal Debates exploratory in SoBigData. More information can be found in www.sobigdata.eu

ESME 2019 - Workshop

Ethical Social Mining and Explainability in AI

8th and 9th July 2019

Officine Garibaldi, Pisa, Italy



THE PRO-RES WORKSHOP ABOUT, ETHICS, PRIVACY AND EXPLAINABLE AI

ESME 2019

8th and 9th July 2019 Officine Garibaldi, Pisa, Italy

Big data analytics and social mining raises a number of ethical issues, especially as companies begin monetizing their data externally for purposes different from those for which the data was initially collected. The scale and ease with which analytics can be conducted today completely changes the ethical framework.

We can now do things that were impossible a few years ago, and existing ethical and legal frameworks cannot prescribe what we should do. Data scientists, data engineers, database administrators and anyone involved in handling big data should have a voice in the ethical discussion about the way data is used.

Moreover Artificial Intelligence is becoming a disruptive technology and resources for innovation are currently dominated by giant tech companies. To ensure European independence and leadership, we must invest wisely by bundling, connecting and opening our AI resources having in mind ethical priorities such as transparency and fairness.

ESME will be a workshop where people from Academia and Industry will openly discuss about these dilemmas in formal and informal sessions.

https://kdd.isti.cnr.it/esme2019

Summer School on Analysing Disinformation

25th - 29th June 2019

Kings College London, UK

Social Media, Online Disinformation, and Elections

GATE general architecture for text engineering

The University Of Sheffield.

The recent past has highlighted the influential role of social networks and other digital media in shaping public debate on current affairs and political issues.

Disinformation and the hyper-partisan media distort societal debates, increase polarisation, and threaten participatory democracy. For instance, the surprising success of Brexit and Trump's election has been, at least partially, attributed to the unprecedented weave of false information that in both cases have polluted online debate before the vote.

Not only does misinformation get significant attention and shares, but also alternative narratives often try to gain credibility through reusing content from mainstream media, often framed so as to undermine reader confidence in the latter.

The aim of this summer school is, firstly, to set out the state-of-the-art and challenges in computational misinformation analysis, followed by lectures and hands-on practical sessions on relevant methods, tools, and datasets.

https://gate-socmedia.group.shef.ac.uk/summer-school-comp-misinfo-analysis-2019

SoBigData Magazine is published under the project N° 654024 | Programme: H2020 | INFRAIA

Duration: 1/09/2015 - 30/08/2019

Editorial Secretariat
info@sobigdata.eu

Editorial Board
Fosca Giannotti
Beatrice Rapisarda
Roberto Trasarti

Layout and Design
Beatrice Rapisarda

Coordinator and Legal representative of the project: Fosca Giannotti | fosca.giannotti@isti.cnr.it

SOBIGDATA News is not for sale but is distributed for purposes of study and research and published online at http://www.sobigdata.eu/newsletter

To subscribe/unsubscribe, please visit http://www.sobigdata.eu/newsletter