Social Mining & Big Data Ecosystem

SoBigData

RESEARCH INFRASTRUCTURE

Project Acronym	***SoBigData***
Project Title	***SoBigData Research Infrastructure*** ***Social Mining & Big Data Ecosystem***
Project Number	***654024***
Deliverable Title	***Evaluation Framework Toolkit and Datasets 1***
Deliverable No.	***D11.2***
Delivery Date	***M30***
Authors	***Roberto Trasarti, Natalia Andrienko, Genevieve Gorrell, Kalina Bontcheva, Luca Pappalardo, Giulio Rossetti, Gianbiagio Curato, Alina Sirbu, Paolo Ferragina, Riccardo Guidotti, Francesca Pratesi, Cristina Muntean, Angelo Facchini, Guido Caldarelli.***

SoBigData receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654024

DOCUMENT INFORMATION

PROJECT	
Project Acronym	SoBigData
Project Title	SoBigData Research Infrastructure Social Mining & Big Data Ecosystem
Project Start	1st September 2015
Project Duration	48 months
Funding	H2020-INFRAIA-2014-2015
Grant Agreement No.	654024
	DOCUMENT
Deliverable No.	D11.2
Deliverable Title	Evaluation Framework Toolkit and Datasets 1
Contractual Delivery Date	February 28 2018
Actual Delivery Date	March 2 2018
Author(s)	Roberto Trasarti, Natalia Andrienko, Genevieve Gorrell, Kalina Bontcheva, Luca Pappalardo, Giulio Rossetti, Gianbiagio Curato, Alina Sirbu, Paolo Ferragina, Riccardo Guidotti, Francesca Pratesi, Cristina Muntean, Angelo Facchini, Guido Caldarelli.
Editor(s)	Roberto Trasarti, Beatrice Rapisarda
Reviewer(s)	Chiara Boldrini, Barbara Furletti
Contributor(s)	
Work Package No.	WP11
Work Package Title	NA5_Evaluation
Work Package Leader	FRH
Work Package Participants	CNR, USFD, UNIPI, FRH, UT, IMT, LUH, KCL, SNS, AALTO, ETHZ, TUDelft
Dissemination	PU
Nature	Report
Version / Revision	V2.0
Draft / Final	Final
Total No. Pages (including cover)	30
Keywords	Evaluation, Exploratory, Methods, Datasets

DISCLAIMER

SoBigData (654024) is a Research and Innovation Action (RIA) funded by the European Commission under the Horizon 2020 research and innovation programme.

SoBigData proposes to create the Social Mining & Big Data Ecosystem: a research infrastructure (RI) providing an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life, as recorded by “big data”. Building on several established national infrastructures, SoBigData will open up new research avenues in multiple research fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, re-use and integration of state-of-the-art big social data, methods, and services, into new research.

This document contains information on SoBigData core activities, findings and outcomes and it may also contain contributions from distinguished experts who contribute as SoBigData Board members. Any reference to content in this document should clearly indicate the authors, source, organisation and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the SoBigData Consortium and its experts, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The European Union (EU) was established in accordance with the Treaty on the European Union (Maastricht). There are currently 27 member states of the European Union. It is based on the European Communities and the member states’ cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice, and the Court of Auditors (http://europa.eu.int/).

Copyright © The SoBigData Consortium 2015. See http://project.sobigdata.eu/ for details on the copyright holders.

For more information on the project, its partners and contributors please see http://project.sobigdata.eu/. You are permitted to copy and distribute verbatim copies of this document containing this copyright notice, but modifying this document is not allowed. You are permitted to copy this document in whole or in part into other documents if you attach the following reference to the copied elements: “Copyright © The SoBigData Consortium 2015.”

The information contained in this document represents the views of the SoBigData Consortium as of the date they are published. The SoBigData Consortium does not guarantee that any information contained herein is error-free, or up to date. THE SoBigData CONSORTIUM MAKES NO WARRANTIES, EXPRESS, IMPLIED, OR STATUTORY, BY PUBLISHING THIS DOCUMENT.

GLOSSARY

ABBREVIATION	DEFINITION
TSMM	Text and Social Media Mining
SNA	Social Network Analysis
HMA	Human Mobility Analytics
WA	Web Analytics
VA	Visual Analytics
SD	Social Data

TABLE OF CONTENT

DELIVERABLE SUMMARY

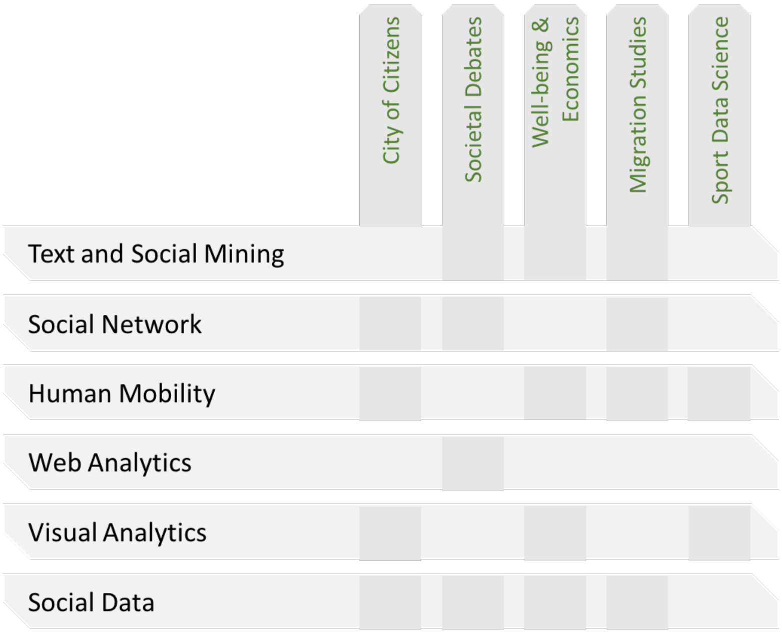
D11.2 contains the SoBigData evaluation data collection toolkit, which will enable the campaign participants to create automatically the evaluation datasets, as described in T11.2. In addition, D11.2 comprises the materials and datasets created for the SoBigData evaluation campaigns, carried out as part of T11.2, and reports on the definition of the exploratories of T11.4. All five thematic areas covered by the SoBigData project have their corresponding datasets: text and social media mining (USFD, UNIPI), social network analysis (CNR, AALTO), human mobility analytics (CNR), web analytics (LUH), visual analytics (FRH). FRH has overall responsibility for coordinating the deliverable production.

1 INTRODUCTION

In this deliverable we report the evaluation of the SoBigData methods, in particular we describe how the quality of the tools provided is measured and assessed. Given the large variety in the methodologies and topics covered very different indicators and measures are used in order to understand the quality of the results obtained. For this reason, we provide for each method a summary description contain the method name, objectives, and its key performance indicators, also reporting the Thematic clusters (listed below) it belongs to (not necessary a single one):

1. Text and Social Media Mining (TSMM)
2. Social Network Analysis (SNA)
3. Human Mobility Analytics (HMA)
4. Web Analytics (WA)
5. Visual Analytics (VA)
6. Social Data (SD)

As already described in D11.1, the SoBigData project relies on vertical thematic environments, called Exploratories, on top of the SoBigData infrastructure, for performing cross-disciplinary social mining research. There are five Exploratories: **City of Citizens**: smart cities, human mobility behavior analysis; **Societal Debates**: text analysis, social network analysis; **Well-Being & Economics**: poverty indicators, spatial analysis; **Migration Studies**: macroscopic human flows, social behavior analysis; **Sport Data Science**: training indicators, performance predictors. As shown in Fig.1 the Thematic Clusters and the Exploratories are two different ways of navigating the methods and the datasets provided by the project.



	City of Citizens	Societal Debates	Well-being & Economics	Migration Studies	Sport Data Science
Text and Social Mining		✓	✓	✓	
Social Network	✓	✓		✓	
Human Mobility	✓		✓	✓	✓
Web Analytics		✓			
Visual Analytics	✓		✓		✓
Social Data	✓	✓	✓	✓	

Figure 1 The Thematic Clusters and Exploratories matrix

In the next sections we survey the methods in the SoBigData platform grouping them by the exploratory they belong to. Cross-Exploratory methods are discussed at the end: they are methods which are general and may be used in different contexts (especially network analysis).

Finally, it is worth noting that not all the methods may be evaluated: for example, some of them are a simple transformation of the data while for others there is no ground truth to be used for the evaluation.

2 CITY OF CITIZENS

This exploratory tells stories about cities and people living in them. Data scientists describe those territories by means of data, statistics and models. This allows citizens and local administrators to better understand cities and how to improve them. Our data scientists study the traffic in cities and their surroundings by analyzing Big Data sources such as mobile phone traces, veicular gps traces and social media data, which are considered as proxies of human behaviour. These results could be useful for both local administrators and citizens. The local administrators have a tool to quantify accurately city traffic and to understand how the city is "used", in order to take better decisions to manage mobility. Citizens could rely on these methods to be informed about the traffic situation in real time so that they could choose the best and fastest way towards their destination.

2.1 MYWAY – TRAJECTORY PREDICTION

Thematic Cluster: HMA
Partners Acronym: SoBigData.it - CNR, KDD
Dataset Used: GPS Tracks - Tuscany
Evaluation: MyWay is a prediction system which exploits the individual systematic behaviors modeled by mobility profiles to predict human movements. MyWay provides three strategies: the individual strategy uses only the user's individual mobility profile, the collective strategy takes advantage of all users' individual systematic behaviors, and the hybrid strategies a combination of the previous two. MyWay only requires the sharing of the individual mobility profiles (a concise representation of the user's movements), instead of raw trajectory data revealing the detailed movements of the users. For the evaluation we considered only the trajectories formed by at least three points, longer than one kilometer and with a duration longer than one minute. Having one month of data, we used the first 3 weeks as training set and the remaining last week as test set. We tested MyWay using two different test sets: the first obtained by considering only the first 33% of each trajectory, and the second by considering the first 66%.
The predictive performance of MyWay is evaluated in terms of accuracy, prediction rate and distance error with respect to the positions predicted and the real one considering a spatiotemporal tolerance. The performance improves drastically when both the individual and collective strategies are used together and when more mobility profiles are shared [1].

2.2 CARPOOLING - CARPOOLING NETWORK ANALYSIS

Thematic Cluster: HMA, SNA
Partners Acronym: SoBigData.it - CNR, KDD
Dataset Used: GPS Tracks - Tuscany
Evaluation: Potential carpooling networks are constructed using mobility data from travelers in a given territory. Nodes correspond to the users and links to the possible shared trips. The structural and topological properties of this network, such as network communities and node ranking, are analyzed to the purpose of highlighting the subpopulations with higher chances to create a carpooling community, and the propensity of users to be either drivers or passengers in a shared car. This study is anchored to reality thanks to the large mobility dataset provided by Octo Telematics, consisting of the complete one-month-long GPS trajectories. We analyze the aggregated outcome of carpooling by means of empirical simulations, showing how an assignment policy exploiting the network analytic concepts of communities and node rankings minimizes the number of single occupancy vehicles observed after carpooling. For the evaluation we considered only the trajectories formed by at least three points, longer than one kilometer and with a duration longer than one minute. We separated working days and non-working days and we filtered out

weekend trajectories. In order to consider also the heterogeneity of the territory we split it into provinces, each containing all the trajectories that pass through it. In particular, we analyzed the results obtained for the Pisa and Florence provinces. The performances show a percentage of single occupancy vehicles as low as 4.63%, which is less than half of what any random assignment can reach. Moreover, as overall result, about 77% of the trips could be saved on the dataset analyzed, and the estimates of saved kms, time, fuel, money and CO2 emissions are significant [2].

2.3 DITRAS - DIARY-BASED TRAJECTORY GENERATOR

Thematic Cluster: HMA
Partners Acronym: SoBigData.it - CNR, KDD
Dataset Used: GPS Tracks - Tuscany
Evaluation: The generation of realistic spatio-temporal trajectories of human mobility is of fundamental importance in a wide range of applications, such as the developing of protocols for mobile ad-hoc networks or what-if analysis in urban ecosystems. Current generative algorithms fail in accurately reproducing the individuals' recurrent schedules and at the same time in accounting for the possibility that individuals may break the routine during periods of variable duration. Ditras (DIary-based TRAjectory Simulator) is a framework for simulating the spatio-temporal patterns of human mobility which operates in two steps: the generation of a mobility diary, and the translation of the mobility diary into a mobility trajectory. We compared the patterns generated by Ditras against real data and synthetic data produced by other generative algorithms. The experimental results show that the proposed algorithm reproduces the statistical properties of real trajectories in the most accurate way, making a step forward in understanding the origin of the spatio-temporal patterns of human mobility [22].

2.4 HUMAN MOBILITY DATA PRIVACY RISK ESTIMATOR

Thematic Cluster: HMA
Partners Acronym: SoBigData.it - CNR, KDD
Dataset Used: GPS Tracks - Tuscany
Evaluation: This method is a fast and flexible approach to estimate privacy risk in human mobility data. The idea is to train classifiers to capture the relation between individual mobility patterns and the level of privacy risk of individuals. We show the effectiveness of our approach by an extensive experiment on real-world GPS data in two urban areas and investigate the relations between human mobility patterns and the privacy risk of individuals [24].

2.5 TRIP BUILDER

Thematic Cluster: HMA
Partners Acronym: SoBigData.it - CNR, HPC
Dataset Used: Flickr and Wikipedia Tourism Trajectories
Evaluation: is a user-friendly and interactive system for planning a time-budgeted sightseeing tour of a city on the basis of the points of interest (PoIs) and the patterns of movements of tourists mined from user-contributed data. The knowledge needed to build the recommendation model is entirely extracted in an unsupervised way from two popular collaborative platforms: Wikipedia and Flickr.

The effectiveness evaluation of Trip Builder is done as such: (i) selecting a set of trajectories of interest for a given user (TRIPCOVER), and (ii) scheduling that set on the user agenda (TRAJSP). This is done by comparing its performance with those obtained by competitive baseline by means of evaluation metrics that consider the actual behavior of test users as mined from Flickr. The evaluation of the efficiency of the Trip Builder framework together covers both TRIPCOVER and TRAJSP solutions.

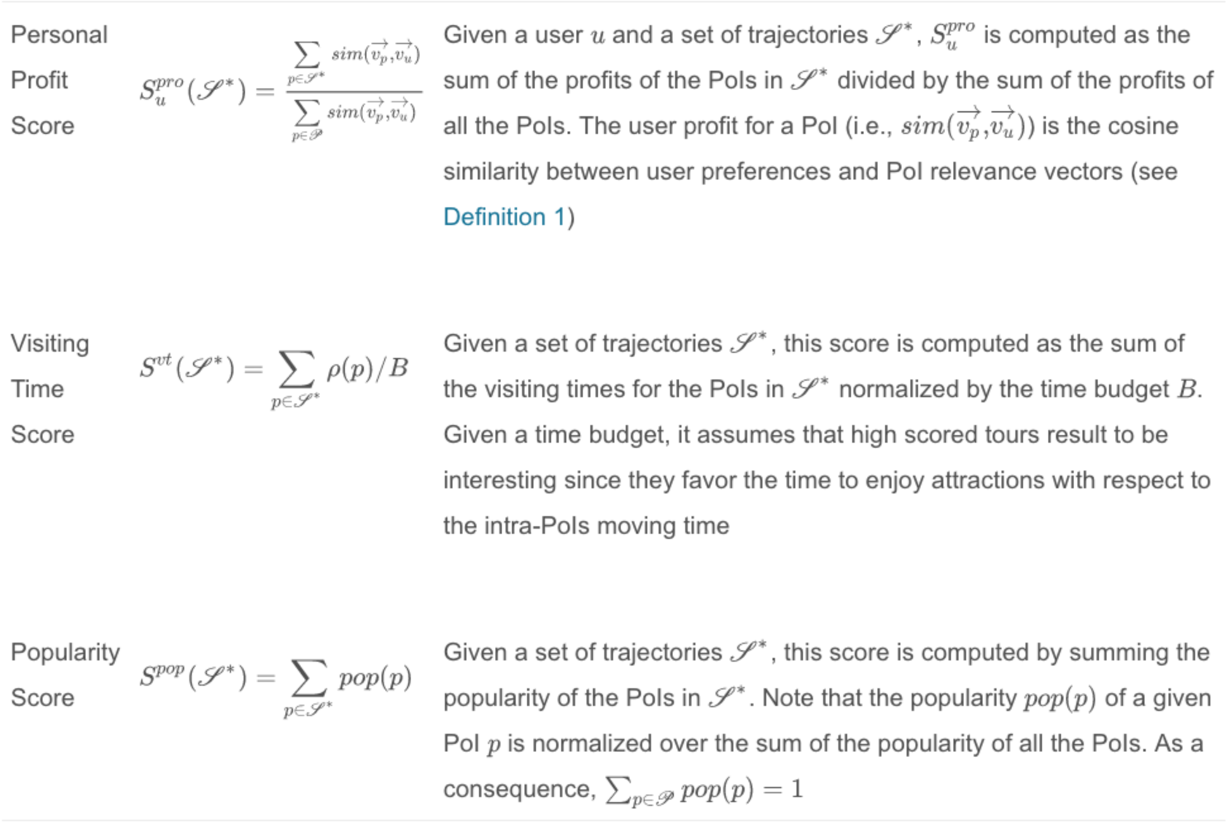
The experiments are conducted on the three datasets of Pisa, Florence, and Rome by varying the time budget and the parameter affecting the contribution of PoIs/user-similarity and PoI-popularity to user profit. Moreover, two different sets of experiments are performed, which differ in the methodology used to choosing the test users:

- Random selection. Here the set of users used to assess Trip Builder performance is randomly chosen. In particular, we consider for all the three cases a set of 100 test users randomly selected among the visitors having a PoI history longer than 10, 15, and 20 PoIs for Pisa, Florence and Rome, respectively. The threshold on the length of the PoI history is set in order to be able to vary in a significant range the time budgets. This is because it is not feasible to evaluate a personalized 4-days itinerary in Rome with test users that actually visited only a few popular PoIs during a single day of visit. By using the above cutoff values, the users among which the 100 test users were chosen are 153, 679, and 930 in Pisa, Florence, and Rome, respectively.
- Profile-based selection. Here we select the test users among users who actually visited at least two of the three cities. In particular, given a user visiting two cities A and B, we used the preference vector obtained from the PoIs visited in city A to generate the personalized sightseeing tour in city B and vice versa. In this way we avoid any possible bias to the specific categories used in the Wikipedia pages of a given city.

Experiments are conducted by providing to Trip Builder and the baseline algorithms the preference vector of each one of the test users in each city, along with a time budget varying in the range 1, 2, and 4 days (1/2, 1 day in the case of the small city of Pisa). We evaluate the performance of the three methods by means of the metrics defined in the following Figure 2. Moreover, we also employ **recall**, a well-known IR metrics that in our settings measures the ability of the methods to predict PoIs and categories that match actual PoI histories of the users in the test set.

The proposed solutions outperform the baselines in terms of all the metrics adopted for assessment. The solution suggests itineraries that better match user preferences. Moreover, such itineraries present higher visiting time and, consequently, lower intra-PoI movement time than the baselines. The tests conducted to demonstrate the efficiency of Trip Builder show that it computes a four-day personalized sightseeing tours of Rome in about 3 seconds thus confirming that the approach can be fruitfully deployed in online applications.

A more detailed view on the entire evaluation process can see found in Brilhante et al., 2015 [26].



Personal Profit Score	$S_u^{pro}(\mathscr{S}^*) = \frac{\sum_{p\in\mathscr{S}^*} sim(\overrightarrow{v_p},\overrightarrow{v_u})}{\sum_{p\in\mathscr{P}} sim(\overrightarrow{v_p},\overrightarrow{v_u})}$	Given a user u and a set of trajectories $\mathscr{S}^*$, S_u^{pro} is computed as the sum of the profits of the PoIs in $\mathscr{S}^*$ divided by the sum of the profits of all the PoIs. The user profit for a PoI (i.e., $sim(\overrightarrow{v_p},\overrightarrow{v_u})$) is the cosine similarity between user preferences and PoI relevance vectors (see Definition 1)
Visiting Time Score	$S^{vt}(\mathscr{S}^*) = \sum_{p\in\mathscr{S}^*} \rho(p)/B$	Given a set of trajectories $\mathscr{S}^*$, this score is computed as the sum of the visiting times for the PoIs in $\mathscr{S}^*$ normalized by the time budget B. Given a time budget, it assumes that high scored tours result to be interesting since they favor the time to enjoy attractions with respect to the intra-PoIs moving time
Popularity Score	$S^{pop}(\mathscr{S}^*) = \sum_{p\in\mathscr{S}^*} pop(p)$	Given a set of trajectories $\mathscr{S}^*$, this score is computed by summing the popularity of the PoIs in $\mathscr{S}^*$. Note that the popularity $pop(p)$ of a given PoI p is normalized over the sum of the popularity of all the PoIs. As a consequence, $\sum_{p\in\mathscr{P}} pop(p) = 1$

Figure 2 Metrics for TripBuilder

2.6 STATISTICALLY VALIDATED NETWORKS

Thematic Cluster: HMA, SNA
Partners Acronym: SNS
Dataset Used: e-MID dataset
Evaluation: This is a theoretical and algorithmic methodology designed to filter out a backbone structure of a complex network by using rigorous statistical testing. It can be applied both to unipartite and to bipartite networks [27][28]. In the bipartite case the method provides a filtering of the projected network, either on the first or on the second module. The filtering is done by statistically comparing the input network with a randomized version of the same network (the Null model), obtained by fixing some properties of the real network (strength/degree distribution) and by letting links to be drawn completely at random once conditioned on the imposed constraints. The statistical filter preserves only the links with a very small p-value in the randomized version of the network. Namely, if one link is very likely to be there (or to have the same or greater weight) in the null model, then the existence of that link (or the size of that weight) is just a statistical consequence of the general structure of the network (i.e. the degree distribution) and not a feature peculiar to that specific network. In [28] the authors compare the trading relationships empirically observed in the e-MID market with a null hypothesis of random trading among banks. They show that the filtering procedure is able to detect preferential trading patterns belonging to the interbank network.

2.7 SOCIOMETER

Thematic Cluster: HMA

Partners Acronym: SoBigData.it - CNR

Dataset Used: CDR Data - Tuscany

Evaluation: The Sociometer is an analytical framework based on data mining methods that analyzes users' call habits, and classifies people into behavioral categories(residents, commuters and visitors). The Sociometer allows to study city users and the impact of big events in cities. The evaluation of this method was carried out in the case of study in Tuscany [35]. Here the data from the Official Statistics, containing the number of residents and dynamic residents (commuters to another area) for each municipality, is compared with the sociometer results. The pearson coefficient is the measure used to study the correlation between the two sources. In particular in the case of study of Tuscany shown in Figure 3, we have a high correlation index and the results are good.

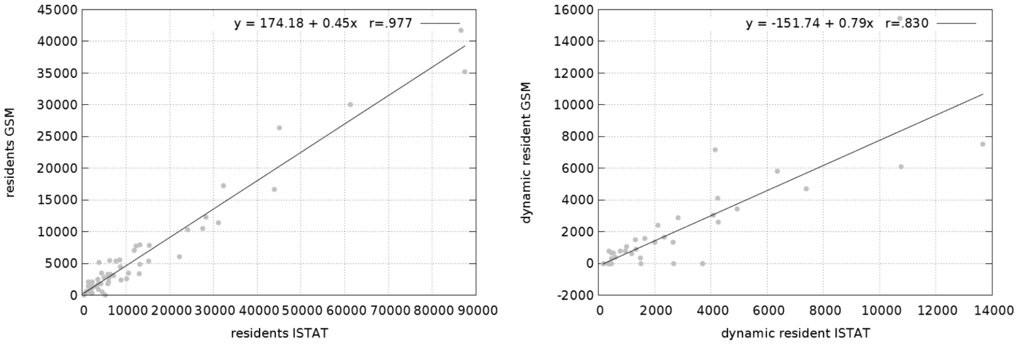


Figure 3 Comparison between Sociometer results and the official statistics

2.8 PRIVACY RISK ON SOCIOMETER

Thematic Cluster: HMA

Partners Acronym: SoBigData.it - CNR

Dataset Used: CDR Data – Tuscany

Evaluation: Given the methodology described in D11.1 for extracting profiles, we can analyze the privacy risks of the users.

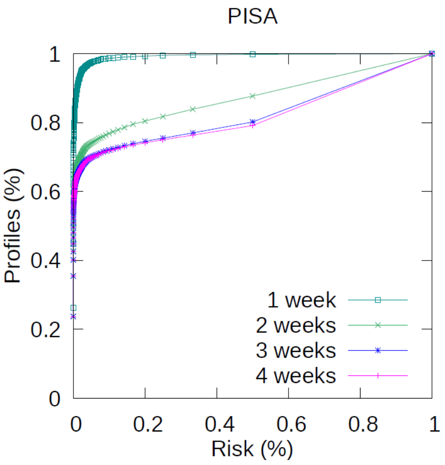


Figure 4 Cumulative distribution of Risk obtained with the attack simulation

The privacy risk in our case is the risk of re-identification, i.e., the probability of an attacker to discover the identity of an individual, having some external information on his target. We assume as background knowledge for the attacks that an attacker knows, for certain municipality, the activities done by a user, in particular the time of all his calls, for a period of 1, 2, 3 or 4 weeks. The simulation of the attacks is performed on profiles built on data collected in November 2015 in Tuscany, for a total of 734,552 users generating 2,121,331 profiles. In Figure 4 we can see the cumulative distribution obtained with the attack simulation, varying the magnitude of the background knowledge, for the municipality of Pisa.

3 SOCIETAL DEBATES

This exploratory is dedicated to the study of public debates in order to discover and understand the most discussed topics. Through the analysis of discussions on social media and articles on newspaper, the methods and tools here collected allow to identify themes, and to track them over space and time.

3.1 TAGME AND WAT: MYWAENTITY DISCOVERY IN TEXTS

Thematic Cluster: TSMM
Partners Acronym: SoBigData.it - UNIPI
Dataset Used: GERBIL
Evaluation: Since 2010 the Acube Lab of UNIPI is studying, designing and implementing Semantic Text Annotators, a.k.a. Entity Linkers. These algorithms are able to detect and annotate sequences of terms with unambiguous and pertinent entities drawn from a catalog (typically, Wikipedia). The result of this effort has been the design of two entity linkers: TagMe [Ferragina-Scaiella, IEEE Software 2012] and WAT [42]. Both algorithms have been refined and engineered in the last six years thus constituting nowadays the best known publicly available annotators in terms of efficiency and efficacy [40]. These tools have been successfully used by their authors in several applications: such as news clustering [ACM WSDM '12] and classification [ECIR '12], analysis of hashtags in tweets [ICWSM '15], and entity salience and relatedness [37,38]. TagMe and WAT are in the SoBigData platform as VREs, for whicha detailed documentation is provided. Our entity linkers have been experimentally evaluated and compared against many others by using the GERBIL dataset and its associated evaluation framework [40] (see also http://aksw.org/Projects/GERBIL.html). The rationale behind this framework is to provide developers, end users and researchers with easy-to-use interfaces that allow for the agile, fine-grained and uniform evaluation of annotation tools on multiple datasets. With the permanent experiment URIs provided by this framework, GERBIL also ensures the reproducibility and archiving of evaluation results, and generates data in machine-processable format thus allowing for the efficient querying and post-processing of evaluation results. Experimental results on the GERBIL platform and dataset have shown that WAT achieves state-of-the-art results on well written texts in terms of F1-measure by approaching other two effective systems, such as PBoH (ETH, 2016) and DoSeR (Passau, 2016), but its annotation speed is about 35 times faster than those ones, thus making WAT useful in large scale applications. TagME is still an interesting entity linker on poorly written texts by achieving more than 70% F1-performance at a very high speed of annotation. Given these properties our two entity annotators got on the SoBigData platform more than 600 millions queries to date.

3.2 BREXIT ANALYZER

Thematic Cluster: TSMM
Partners Acronym: USFD
Dataset Used: Brexit Twitter User Vote Intent, UK General Election Vote Intent
Evaluation: Classification of users according to referendum vote intent is a central part of the Brexit Analyzer Pipeline, and was done on the basis of tweets authored by them and identified as being in favour of leaving or remaining in the EU. Such tweets were identified based on 59 hashtags indicating allegiance. Hashtags in the final position more reliably summarise the tweeter's position, so only these were used. Consider, for example, "is Britain really #strongerin? I don't think so! #voteleave". This approach was evaluated using a set of users that explicitly declared their vote intent in response to Brndstr's Twitter campaign offering a topical profile image modification. The formulaic tweet required to obtain the image modification enabled a ground truth sample to be easily and accurately gathered. On these data, we found our method produced a 94% accuracy even on the basis of a single partisan tweet (where three are required, an accuracy of 99% can be obtained, though only 60,000 such users can be found, as opposed to 290,000 with at least one partisan tweet). The Brndstr data itself was also used to supplement the set, raising the accuracy further, and resulting in a list of 208,113 leave voters and 270,246 remain voters.

A further key piece of information is the accurate classification of users according to the political party they support. Hashtags were again used to automatically identify party supporters. As for Brexit vote intent, tweets with such hashtags in the final position were used to identify party supporters. Additionally, a further method considered party allegiance expressed in the Twitter biography. Biographies were used to provide information for manual annotation of the gold standard set, even in the case of users classified on the basis of hashtags, because bios provide a more informative short passage than a tweet, which may not provide enough information in and of itself. On a sample of 220 bios that were double-annotated, a three-way interannotator agreement of 0.961 was achieved. Thereafter, a single annotator was considered sufficient for the remainder of the sample. A total of 909 users were annotated. Overall, the approach has an accuracy of 0.93 (bios method) and 0.99 (hashtags method). Factoring in the actual proportions of users classified using each method, for each party, in our final lists, the accuracy of the system is estimated to be around 0.97. In total, 73,500 users have been classified, of whom 57,000 were Labour supporters, 9,000 were SNP supporters, 4,500 were Liberal Democrat supporters, 2,500 were Conservative supporters and the remainder belonged to smaller parties.

4 WELL-BEING AND ECONOMY

This exploratory investigates the changes in the behaviour of people and companies as a consequence of the economic crisis. The measurement of the real cost of life is investigated by studying the price variation, and the correlation of people well-being with their social and mobility data is studied. This exploratory contains approaches that can potentially contribute to the development of effective policies in order to reduce internal and external risks for companies, which can result in systematic improvement of well-being.

4.1 MAXANDSAM NETWORK RECONSTRUCTION METHOD

Thematic Cluster: SNA
Partners Acronym: SoBigData.it - IMT
Dataset Used: e-MID dataset, e-MID interbank transactions
Evaluation: this method aims at reconstructing economic and financial networks, taking as input nodes fluxes (e.g. assets and liabilities, exports and imports) as well as the total number of observed links. The latter define the probability for any two banks to have a transaction, as well as the expected magnitude of the transaction itself. The method has been recently extended to implement the reconstruction of bipartite networks too [3,4]. The reconstruction provided by our method has been compared with the performance of other similar algorithms. Remarkably, these "horseraces" have highlighted that our method is "the clear winner" among the ensemble algorithms [5,6]. The measures used for the evaluation of these methods are "structural" in nature, i.e. they concern quantities of interest for the reconstruction of the network topology/weights (accuracy, Jaccard similarity, Hamming distance, Cosine similarity, Shannon-Jensen divergence, scatter plots of observed VS expected quantities, empirical CDFs).

4.2 DEBTRANK SYSTEMIC RISK ESTIMATION METHOD

Thematic Cluster: SNA
Partners Acronym: SoBigData.it - IMT
Dataset Used: e-MID dataset, e-MID interbank transactions
Evaluation: this method aims at providing a measure of distress of financial institutions. DebtRank is an iterative method quantifying the impact of subsequent (financial) shockwaves on the entities constituting the network under analysis. It complements the usual way of running stress tests - which consider defaulted institutions only - by quantifying their "closeness" to default. From a purely structural point of view, it implements the "too-connected-to-fail" concept instead of the more popular "too-big-to-fail" [7]. Although DebtRank represents just one out of many possible indicators of risk [3,7], it has recently gained increasing attention, being employed by the ECB to monitor TARGET2 [8]. It has been also tested in combination with the MaxAndSam reconstruction method, being accurately reproduced.

4.3 MAXIMUM-ENTROPY NETWORK RECONSTRUCTION

Thematic Cluster: SNA
Partners Acronym: SNS
Dataset Used: FED data
Evaluation: The methodology reconstructs bipartite networks from the knowledge of nodes' strengths only, via maximization of the entropy function. An application to systemic risk analysis is presented in [29]. The rationale behind the use of maximum entropy is that it enables the reconstruction of the (bipartite) network of portfolio compositions of companies by only knowing (publicly available) node features, namely size and leverage of each company and total capitalization of each asset class. In [29] it is shown that the systemic risk measures introduced in [30], i.e. aggregated systemicness (the percentage of aggregate equity wiped out as a consequence of a negative asset class shock) and systemicness of single banks (the

contribution of a single bank to the aggregate systemicness), calculated on the reconstructed network is a good approximation of the same metric calculated on the real network (of credits and liabilities). Thus, the method allows for systemic risk assessment from partial information.

4.4 NETWORK CONSTRUCTION VIA TAIL GRANGER-CAUSALITY

Thematic Cluster: SNA
Partners Acronym: SNS
Dataset Used: bond yield, equity log-returns and CDS spreads
Evaluation: Given a set of time series, the methodology builds a network by inferring causality of rare-events. The adopted method is Granger-causality in tails [32], i.e. it is tested whether an extreme events in one time series helps predicting the occurrence of a future extreme event in another time series. This method was applied in [31] to construct a bipartite network of systemically important banks and sovereign bonds, where the presence of a link between two nodes indicates the existence of a Granger tail causal relation. This means that tail events in the equity variation of a bank helps in forecasting a tail event in the price variation of a bond, i.e. forecast episodes of systemic risk. An out of sample analysis shows that connectedness and centrality network metrics, e.g. the degree of bond nodes, have a significant cross-sectional forecasting power of bond quality measures.

5 MIGRATION STUDIES

Could Big Data help to understand the migration phenomenon? In this exploratory we try to answer various questions about migration in Europe and in the world. Several studies are ongoing, including developing economic models of migration, nowcasting migration stocks and flows, identifying perception of migration and effect on the leaving and the receiving communities. At the moment this exploratory is still in population and will be released soon in the platform (M49)

5.1 NEXT INSTITUTION PREDICTION BASED ON SCIENTIFIC PROFILE

Thematic Cluster: HMA
Partners Acronym: SoBigData.it - CNR
Dataset Used: Scientific Publications Dataset
Evaluation: This method aims at predicting the future institution of a scientist given her recent scientific profile [36]. In the first phase, a data mining approach is used to predict whether or not a scientist will migrate, obtaining an AUC=0.85 (significantly better than a baseline method having AUC=0.50). In the second phase, the next institution is predicted by using a social-gravity model, which produces an error in the prediction which is 85% lower than using the traditional gravity model.

5.2 EPIDEMIC SENTIMENT ANALYSIS

Thematic Cluster: TSMM
Partners Acronym: SoBigData.it - UNIPI, CNR
Dataset Used: Twitter Stream Dataset, Semeval2013, Semeval2014, Earth Hour 2015
Evaluation: This is a method based on epidemic spreading to automatically extend the dictionary used in lexicon-based sentiment analysis, starting from a reduced dictionary and large amounts of Twitter data [43]. We evaluate the method by computing correlation between the new dictionary and a manually annotated one (test dictionary). The resulting dictionary is shown to contain valences that correlate well with human-annotated sentiment, with values of 0.7. Secondly, we compare the sentiment classifications obtained by our method with those from the Semeval and Earth Hour datasets, and we see results comparable to the original dictionary in terms of accuracy, recall, precision and F1-values. However we are able to tag more tweets compared to the original.

6 SPORT DATA SCIENCE

This exploratory tells stories about sports analytics. Sports data scientists describe performances by means of data, statistics and models. This allows coaches, fans and practitioners to better understand and boost sports performance. Our data scientists are using massive data describing several sports – especially soccer, cycling and rugby – to construct interpretable and easy-to-use tool for sports coaches and managers. These studies open an interesting perspective on how to understand the factors influencing sports success and how to build simulation tools for boosting both individual and collective performance. At the moment this exploratory is still in population and will be released soon in the platform (M49)

6.1 SOCCER TEAMS RANKING SIMULATOR

Thematic Cluster: HMA
Partners Acronym: SoBigData.it - Unipi
Dataset Used: Soccer Team Performance
Evaluation: This simulator produces a ranking of soccer teams in a league, on the basis of their technical performances during a season. In particular, for each game in a season the simulator generates a synthetic outcome only relying on technical data, i.e., excluding the goals scored, exploiting an outcome predictor trained on data from past seasons. We validated the simulator by using more than 6,000 games and 10 million events in six European leagues. The simulation produces a team synthetic ranking which is similar to the actual ranking, suggesting that a complex systems' view on soccer has the potential of revealing hidden patterns regarding the relation between performance and success [23].

7 CROSS-EXPLORATORIES METHODS

In the following we summarise the cross-exploratory methods, i.e., the general-purpose methods that have been used in different exploratories.

7.1 SWAT MYWAENTITY SALIENCE IN TEXTS

Thematic Cluster: TSMM
Partners Acronym: SoBigData.it - Unipi
Dataset Used: NewYork Times (payment needed) and Wikinews
Evaluation: SWAT is a software system that solves efficiently and effectively the document aboutness problem by providing a succinct representation of a document's subject matter via salient entities drawn from Wikipedia. At the time of SWAT proposal [37], the literature offered two systems: the Cmu-Google system, which used a proprietary entity annotator to extract entities from the input text and a very simple binary classifier based on very few and basic features to distinguish between salient and non-salient entities, and the SEL system that hinged on a supervised two-step algorithm comprehensively addressing both entity annotation and entity-salience scoring. Our system SWAT introduces three main novelties: (i) it carefully orchestrates state-of-the-art tools, publicly available in IR and NLP literature, to extract several new features from the syntactic and the semantic elements of the input document which are suitable for establishing the salience of entities; (ii) it builds a binary classifier based on these features that achieves improved micro- and macro-F1 performance; (iii) it is released to the community in order to allow its use as a module within other tools. The experimental evaluation of SWAT has been executed over two datasets which are very well known for this problem and have the following features. The annotated version of NewYork Times (NYT), suitable for the document aboutness problem, which was introduced in 2014 and consists of annotated news drawn from 20 years of the NYT newspaper for a total of about 110k news and 1.4 million of annotated entities; and the Wikinews dataset, which was introduced in 2016 and consists of a sample of 365 news published by Wikinews from November 2004 to June 2014 and annotated with about 5000 entities by the Wikinews community. Although the latter dataset is significantly smaller than NYT, it has some remarkable features with respect to NYT: the ground-truth generation of the salient entities was obtained via human-assigned scores rather than being derived in a rule-based way, and it includes both proper nouns (as in NYT) and common nouns (unlike NYT) as salient entities. Our experiments have shown that SWAT raises the known state-of-the-art performance of the previously known systems in terms of F1 up to about 11% (absolute) over Cmu-Google system and up to 5% (absolute) over SEL. These F1-results have been complemented with a throughout study of the contribution of each feature (old and new ones) and an evaluation of the performance of known systems in dealing with documents where salient entities are not necessarily biased to occur at their beginning. In this specific setting, experiments have shown that the improvement of SWAT with respect to Cmu-Google over the largest dataset NYT gets up to 14% in micro-F1.

7.2 SMAPH: MYWAENTITY DISCOVERY IN QUERIES

Thematic Cluster: TSMM
Partners Acronym: SoBigData.it - Unipi
Dataset Used: GERDAQ
Evaluation: SMAPH is a software system that realizes the linking of open-domain web-search queries towards entities drawn from Wikipedia [39,41]. It is a second-order approach that, by piggybacking on a web search engine (either Bing or Google, in the following experiments), alleviates the noise and irregularities that characterize the language of queries and puts queries in a larger context in which it is easier to make sense of them. The key algorithmic idea underlying SMAPH is to first discover a candidate set of entities and then link-back those entities to their mentions occurring in the input query. This allows us to confine the possible concepts pertinent to the query to only the ones really mentioned in it. The link-

back is implemented via a collective disambiguation step based upon a supervised ranking model that makes one joint prediction for the annotation of the complete query optimizing directly the F1 measure. We have demonstrated, via a systematic and throughout set of experiments, that SMAPH achieves state-of-the-art performance on the ERD@SIGIR2014 benchmark and on the GERDAQ dataset, the latter has been constructed by us and includes 1000 well-curated queries that have been labeled via a two-phase crowdsourcing process. The experimental results showed that: (i) in the detection of Named Entities, SMAPH is 12% better in macro-F1 than WAT, which is in turn better than other known entity linkers (such as AIDA); (ii) in the detection of generic entities, SMAPH is again the best annotator in terms of macro-F1, achieving an absolute improvement of 12.7% (when Bing is used as piggy-back search engine) and 16.3% (using Google) over WAT; (iii) in the detection of entities and their mentions in queries, again, off-the-shelf entity linkers (such as AIDA and WAT) are worse than SMAPH of about 11%–17% in macro-F1 (using both Google and Bing). As far as other entity-linkers in queries are concerned, since they are not available to the public, the only experimental comparison available is the one performed at the ERD 2014 Short Track Challenge (ACM SIGIR 2014) in which SMAPH was the top-1 and won the competition.

7.3 DEMON

Thematic Cluster: SNA
Partners Acronym: SoBigData.it - CNR, UNIPI, KDD
Dataset Used: IMDB Network, Amazon Network, Congress Network
Evaluation: DEMON is a bottom-up node-centric community discovery algorithm [9, 10]. We evaluated the performances of DEMON by comparing the obtained network clusters to the ones produced by state-of-art competitors in terms of: (i) community size and overlap distribution, (ii) interpretability of identified clusters, (iii) ability to retrieve external ground truth partitioning. Moreover, partition quality evaluation was performed using a BLR classifier and a cohesion index. DEMON was applied to address several analytical tasks, among them: support to homophily and service-usage analysis [11, 12], support to link prediction in dynamic networks [13, 14], support to network quantification analysis [15], analysis of mobility functional areas [16]. In all the analyzed scenarios the partitions extracted using the proposed approach lead to the best observed performances w.r.t. the compared competitors.

7.4 TILES

Thematic Cluster: SNA
Partners Acronym: SoBigData.it - CNR, UNIPI, KDD
Dataset Used: Amazon Network, Social Network dataset - LiveJournal, Facebook wallpost, WEIBO interactions
Evaluation: TILES is a bottom-up node-centric community discovery algorithm designed for time evolving networks [17]. We evaluated the performances of TILES by comparing the obtained network clusters to the ones produced by state-of-art competitors in terms of: (i) community size and overlap distribution, (ii) interpretability of identified clusters, (iii) execution time, (iv) ability to retrieve external ground truth partitioning (in terms of Normalised Mutual Information - NMI - a measure that evaluate the adherence an identified partitioning to the expected one).

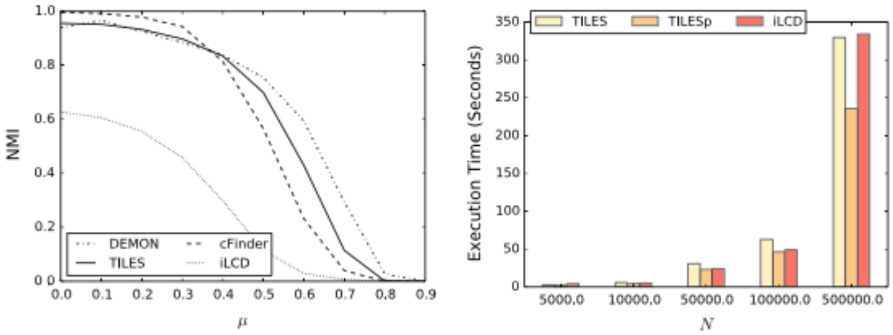


Figure 5 Dynamic Community Discovery: (left) NMI comparison, (right) TILES execution time

Our results, highlighted in the Fig. 5, underline that the proposed method is always able to outperform its direct competitor (iLCD, the stat of art approach for dynamic community discovery in graph streams) both in terms of NMI score as well as in terms of execution time while producing results having comparable quality w.r.t. methods designed for static community discovery (DEMON, cFinder).

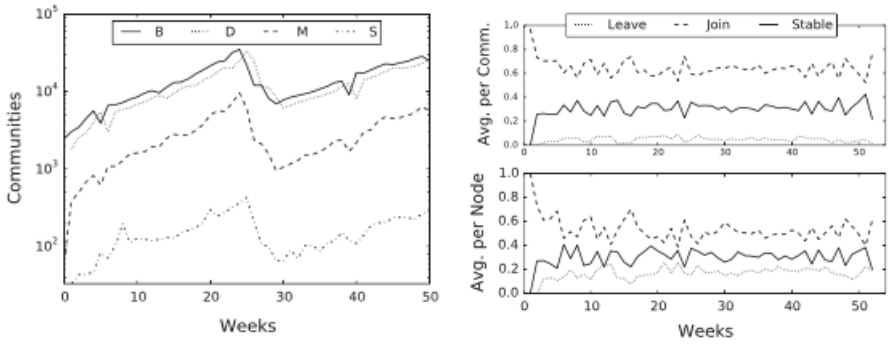


Figure 6 TILES Community Stability: (left) trends for community events - Birth, Merge, Split, Death; (right) example of trends for node/community stability

Moreover, an event based analysis of community life-cycle was performed in order to characterize the identified evolving substructures and their temporal-stability (i.e., the degree of stability the node partition maintains as the underlying network topology changes). Using data from a chinese Twitter-like platform (Sina Weibo, results shown in Fig. 6) we tracked the community events expressed by the evolving topology of online interactions among its users. We observed that, tuning TILES parameters, we can identify the temporal granularity to use in order to describe community evolution as as stable process reducing the impacts of sudden volatile events often related to noisy data.

7.5 NDLIB/NDLIB-REST

Thematic Cluster: SNA
Partners Acronym: SoBigData.it - CNR, UNIPI, KDD
Dataset Used: Synthetic graph generators
Evaluation: NDlib is a python library that allows to easily describe network diffusion simulations [17, 18]. NDlib allows to evaluate and compare different algorithmic models employing standard trend visualisation plots. So far, it was used to introduce novel models [20] as well as to compare existing ones in heterogeneous network settings [21]. To evaluate the library we compared it to various similar libraries in the literature. Qualitatively, we looked at the various features, and showed that our library implements a

more complete set compared to the others. Quantitatively, we compared running times and scalability on the SIR model, and showed that our framework is one order of magnitude faster for various network sizes.

7.6 EGONETWORKS

Thematic Cluster: SNA
Partners Acronym: CNR
Dataset Used: Facebook EuroSys 2009, Facebook - New Orleans regional network
Evaluation: This python package contains classes and functions for the structural analysis of ego networks. An ego network is a simple model that represents a social network from the point of view of an individual. This model considers only the social relationships that a focal node in the network (termed ego) maintains with other nodes (termed alters). Note that the model supported by this package does not consider relationships between alters (a.k.a. mutual friendship relationships), but only the star topology of alters connected to the ego. This ego network model is known as “Dunbar’s ego network”. See [33] and [34] for additional information about ego networks and ego network analysis. The package offers several methods for the static and dynamic analysis of ego networks. For example, the package provides a function to obtain the “social circles” of the ego network, which are discrete groups of alters at similar level of tie strength with the ego. In addition, there are functions to analyse the dynamic evolution of ego networks and to calculate their stability over time. These functions are useful, for example, for the analysis of human behaviour in different social environments as well as to identify particularly active, dynamic or sociable people from their communication traces. The package offers specialised classes for building and studying ego networks from Twitter data and from coauthorship or collaboration networks (i.e. networks where the ego is an author and the alters are people with whom he or she co-authored publications). As far as the evaluation of this method is concerned, to the best of our knowledge this is the only publicly available package that handles the computation of Dunbar’s ego networks features, such as active network size, optimal number of circles, and circles size.

8 CONCLUSIONS

In this deliverable we reported a series of methods available on the SoBigData platform, we have summarised their evaluation criteria, and we have referenced the datasets in the SoBigData catalogue that can be used to run these methods. This represents the SoBigData standard process of method evaluation, granting the high quality of the tools provided by the project. Moreover, this evaluation is replicable according to the constraints of the datasets (in Virtual or Transnational Access), allowing the final users to compare results of their own processes.

REFERENCES

[1] Trasarti, R., Guidotti, R., Monreale, A., & Giannotti, F. (2017). Myway: Location prediction via mobility profiling. Information Systems, 64, 350-367.

[2] Guidotti, R., Nanni, M., Rinzivillo, S., Pedreschi, D., & Giannotti, F. (2017). Never drive alone: Boosting carpooling with network analysis. Information Systems, 64, 237-257. ISO 690

[3] Cimini, G., Squartini, T., Garlaschelli, D., & Gabrielli, A. (2015). Systemic Risk Analysis on Reconstructed Economic and Financial Networks. Scientific Reports 5, 15758, doi:10.1038/srep15758

[4] Squartini, T., Almog, A., Caldarelli, G., van Lelyveld, I., Garlaschelli, D., & Cimini, G. (2017). Enhanced capital-asset pricing model for the reconstruction of bipartite financial networks. Physical Review E 96, 032315

[5] Anand, K., et al. (2017). The missing links: A global study on uncovering financial network structures from partial data. Journal of Financial Stability, doi:10.1016/j.jfs.2017.05.012

[6] Mazzarisi, P., & Lillo, F. (2017). Methods for Reconstructing Interbank Networks from Limited Information: A Comparison. Econophysics and Sociophysics: Recent Progress and Future Directions, 201-215, Springer International Publishing, doi:10.1007/978-3-319-47705-3_15

[7] Bardoscia, M., Battiston, S., Caccioli, F., & Caldarelli, G. (2015). DebtRank: a microscopic foundation for shock propagation. PLoS ONE 10(6): e0130406, doi:10.1371/journal.pone.0130406

[8] https://www.ecb.europa.eu/events/pdf/conferences/140623/2014-06-23_Presentation_of_MaRs_report_at_the_concluding_MaRs_conference_by_P_Hartmann_rev.pdf?965da4986299fc406b7c1418e5a17d1d

[9] Coscia, Michele; Rossetti, Giulio; Giannotti, Fosca; Pedreschi, Dino "DEMON: a Local-First Discovery Method for Overlapping Communities" SIGKDD international conference on knowledge discovery and data mining, pp. 615-623, IEEE ACM, 2012, ISBN: 978-1-4503-1462-6.

[10] Coscia, Michele; Rossetti, Giulio; Giannotti, Fosca; Pedreschi, Dino "Uncovering Hierarchical and Overlapping Communities with a Local-First Approach" ACM Transactions on Knowledge Discovery from Data (TKDD), 9 (1), 2014.

[11] Rossetti, Giulio; Pappalardo, Luca; Kikas, Riivo; Pedreschi, Dino; Giannotti, Fosca; Dumas, Marlon "Community-centric analysis of user engagement in Skype social network" International conference on Advances in Social Network Analysis and Mining , pp. 547-552, IEEE, 2015, ISBN: 978-1-4503-3854-7.

[12] Rossetti, Giulio; Pappalardo, Luca; Kikas, Riivo; Pedreschi, Dino; Giannotti, Fosca; Dumas, Marlon "Homophilic network decomposition: a community-centric analysis of online social services" Social Network Analysis and Mining, 2016.

[13] Rossetti, Giulio; Guidotti, Riccardo; Pennacchioli, Diego; Pedreschi, Dino; Giannotti, Fosca "Interaction Prediction in Dynamic Networks exploiting Community Discovery" International conference on Advances in Social Network Analysis and Mining, pp. 553-558 , IEEE, 2015, ISBN: 978-1-4503-3854-7 .

[14] Rossetti, Giulio; Guidotti, Riccardo; Miliou, Ioanna; Pedreschi, Dino; Giannotti, Fosca "A Supervised Approach for Intra-/Inter-Community Interaction Prediction in Dynamic Social Networks" Social Network Analysis and Mining, 2016.

[15] Milli, Letizia; Monreale, Anna; Rossetti, Giulio; Pedreschi, Dino; Giannotti, Fosca; Sebastiani, Fabrizio "Quantification in Social Networks" International Conference on Data Science and Advanced Analytics, IEEE, 2015.

[16] Gabrielli, Lorenzo; Fadda, Daniele; Rossetti, Giulio; Nanni, Mirco; Piccini, Leonardo; Giannotti, Fosca; Pedreschi, Dino; Lattarulo, Patrizia "Discovering Mobility Functional Areas: A Mobility Data Analysis Approach" 9th Conference on Complex Networks, CompleNet, Forthcoming.

[17] Rossetti, Giulio; Pappalardo, Luca; Pedreschi, Dino; Giannotti, Fosca "Tiles: an online algorithm for community discovery in dynamic social networks" Machine Learning Journal, 2016.

[18] Rossetti, Giulio; Milli, Letizia; Rinzivillo, Salvatore; Sirbu, Alina; Pedreschi, Dino; Giannotti, Fosca "NDlib: a Python Library to Model and Analyze Diffusion Processes Over Complex Networks" International Journal of Data Science and Analytics, 2017.

[19] Rossetti, Giulio; Milli, Letizia; Rinzivillo, Salvatore; Sirbu, Alina; Pedreschi, Dino; Giannotti, Fosca "NDlib: Studying Network Diffusion Dynamics Inproceedings" IEEE International Conference on Data Science and Advanced Analytics, Forthcoming.

[20] Milli, Letizia; Rossetti, Giulio; Pedreschi, Dino; Giannotti, Fosca "Information Diffusion in Complex Networks: The Active/Passive Conundrum" Complex Networks, 2017.

[21] Letizia, Milli; Giulio, Rossetti; Dino, Pedreschi; Fosca, Giannotti "Diffusive Phenomena in Dynamic Networks: a data-driven study" 9th Conference on Complex Networks, CompleNet, Forthcoming.

[22] Pappalardo, Luca; Simini, Filippo "Data-driven generation of spatio-temporal routines in human mobility" Data Mining and Knowledge Discovery, doi:10.1007/s10618-017-0548-4, 2017.

[23] Pappalardo, Luca; Cintia, Paolo "Quantifying the relation between performance and success in soccer" Advances in Complex Systems, doi:10.1142/S021952591750014X, 2017.

[24] Pellungrini, Roberto; Pappalardo, Luca; Pratesi, Francesca; Monreale, Anna "A data mining approach to estimate privacy risk in human mobility data" ACM Transactions on Intelligent Systems and Technology (TIST), 9(3), pp. 31:1–31:27, doi:10.1145/3106774, 2018.

[25] Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., & Zubiaga, A. (2017). SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. arXiv preprint arXiv:1704.05972.

[26] Igo Ramalho Brilhante, Jose Antonio Macedo, Franco Maria Nardini, Raffaele Perego, Chiara Renso, On planning sightseeing tours with TripBuilder, Information Processing & Management, Volume 51, Issue 2, 2015, Pages 1-15, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2014.10.003.

[27] Tumminello, M., Micciche, S., Lillo, F., Piilo, J., & Mantegna, R. N. (2011). Statistically validated networks in bipartite complex systems. *PloS one*, *6*(3), e17994.

[28] Hatzopoulos, V., Iori, G., Mantegna, R. N., Micciché, S., & Tumminello, M. (2015). Quantifying preferential trading in the e-MID interbank market. *Quantitative Finance*, *15*(4), 693-710.

[29] Di Gangi, Domenico and Lillo, Fabrizio and Pirino, Davide, Assessing Systemic Risk Due to Fire Sales Spillover Through Maximum Entropy Network Reconstruction (January 15, 2018). Available at SSRN: https://ssrn.com/abstract=2639178 or http://dx.doi.org/10.2139/ssrn.2639178 .

[30] Greenwood, R., Landier, A., & Thesmar, D. (2015). Vulnerable banks. *Journal of Financial Economics*, *115*(3), 471-485.

[31] Corsi, Fulvio and Lillo, Fabrizio and Pirino, Davide, Measuring Flight-to-Quality with Granger-Causality Tail Risk Networks (March 10, 2015). Available at SSRN: https://ssrn.com/abstract=2576078 or http://dx.doi.org/10.2139/ssrn.2576078 .

[32] Hong, Y., Liu, Y., & Wang, S. (2009). Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics*, *150*(2), 271-287.

[33] R.I.M. Dunbar, V. Arnaboldi, M. Conti, A. Passarella, “The Structure of Online Social Networks Mirrors Those in the Offline World”, Social Networks, Vol. 43, October 2015, Pages 39-47

[34] Valerio, A. Passarella, M. Conti, R.I.M. Dunbar, “Online Social Networks: Human Cognitive Constraints in Facebook and Twitter Personal Graphs”, A volume in Computer Science Reviews and Trends, Elsevier, ISBN: 978-0-12-803023-3, 2015

[35] Furletti B., Gabrielli L.,Garofalo G., Giannotti F., Milli M., Nanni M., Pedreschi D., Vivio R.. Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach. 47th SIS Scientific Meeting of the Italian Statistica Society

[36] James C., Pappalardo L., Sirbu A., Simini F., Prediction of next career moves from scientific profile, eprint arXiv:1802.04830, 2018.

[37] Ponza, M., Ferragina, P., Piccinno, F.. Document Aboutness via Sophisticated Syntactic and Semantic Features. NLDB 2017: 441-453.

[38] Ponza, M., Ferragina, P., Chakrabarti, S.. A Two-Stage Framework for Computing Entity Relatedness in Wikipedia. ACM CIKM 2017: 1867-1876.

[39] Cornolti, M., Ferragina, P., Ciaramita, M., Rüd, S., Schütze, H.. A Piggyback System for Joint Entity Mention Detection and Linking in Web Queries. WWW 2016: 567-578

[40] Usbeck, R., Röder, M., et alii. GERBIL: General Entity Annotator Benchmarking Framework. WWW 2015: 1133-1143

[41] Cornolti, M., Ferragina, P., Ciaramita, M., Rüd, S., Schütze, H.. The SMAPH system for query entity recognition and disambiguation. ERD@SIGIR 2014: 25-30.

[42] Piccinno, F., Ferragina, P.. From TagME to WAT: a new entity annotator. ERD@SIGIR 2014: 55-62.

[43] Pollacci, L., Sîrbu, A., Giannotti, F., Pedreschi, D., Lucchese, C. and Muntean, C.I., 2017, November. Sentiment Spreading: An Epidemic Model for Lexicon-Based Sentiment Analysis on Twitter. In Conference of the Italian Association for Artificial Intelligence (pp. 114-127). Springer, Cham.