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 2***
Deliverable No.	**D11.3**
Delivery Date	***December 2019***
Authors	**Genevieve Gorrell (USFD), Kalina Bontcheva (USFD)**

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.	D11.3
Deliverable Title	Evaluation Framework Toolkit and Datasets 2
Contractual Delivery Date	November 2019
Actual Delivery Date	12 Dicember 2019
Author(s)	Genevieve Gorrell (USFD), Kalina Bontcheva (USFD)
Editor(s)	Genevieve Gorrell (USFD), Beatrice Rapisarda (CNR)
Reviewer(s)	Rajesh Sharma (UT), Chiara Boldrini (CNR)
Contributor(s)	Roberto Trasarti (CNR), Natalia Andrienko (FRH), Luca Pappalardo (CNR), Giulio Rossetti (CNR), Gianbiagio Curato (SNS), Alina Sirbu (UNIPI), Paolo Ferragina (UNIPI), Riccardo Guidotti (CNR), Francesca Pratesi (CNR), Cristina Muntean (CNR), Angelo Facchini (IMT), Guido Caldarelli (IMT), Tiziano Squartini (IMT)
Work Package No.	WP11
Work Package Title	NA5_Evaluation
Work Package Leader	USFD
Work Package Participants	USFD, UNIPI, CNR, AALTO, LUH, FRH
Dissemination	PU
Nature	Report
Version / Revision	V1.2
Draft / Final	Final
Total No. Pages (including cover)	42
Keywords	Evaluation, resources, methods

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/).

GLOSSARY

ABBREVIATION	DEFINITION
CNN	A convolutional neural network (CNN, or ConvNet) is a class of deep neural networks.
ELMo	A deep contextualized word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy).
EU	European Union
F1	F_1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score.
GATE	General Architecture for Text Engineering, NLP technology from USFD.
LSTM	Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning.
NERD	Named entity recognition and disambiguation.
NLP	Natural Language Processing.
RMSE	Root Mean Squared Error, an evaluation metric based on deviation from the correct answer. Lower is better.

TABLE OF CONTENT

DOCUMENT INFORMATION **2**

DISCLAIMER **3**

GLOSSARY **4**

TABLE OF CONTENT **5**

DELIVERABLE SUMMARY **7**

EXECUTIVE SUMMARY **8**

1 Relevance to SoBigData **9**

1.1 PURPOSE OF THIS DOCUMENT 9

1.2 RELEVANCE TO PROJECT OBJECTIVES 9

1.3 SOBIGDATA PROJECT DESCRIPTION 9

1.4 RELATION TO OTHER WORKPACKAGES 9

1.5 STRUCTURE OF THE DOCUMENT 10

2 Supervised EvaluatioN **11**

2.1 Use cases **11**

2.1.1 MYWAY - TRAJECTORY PREDICTION 11

2.1.2 Human mobility data Privacy risk Estimator 11

2.1.3 Statistically Validated networks 12

2.1.4 Machine learning prediction of volatility in bitcoin 12

2.1.5 Machine learning prediction of volatility in bitcoin 13

2.1.6 Sociometer 14

2.1.7 Privacy Risk on Sociometer 15

2.1.8 Next Institution Prediction based on Scientific Profile 15

2.1.9 Epidemic sentiment analysis 16

2.1.10 Superdiversity and Sentiment 17

2.1.11 SWAT: MyWaEntity salience in texts 17

2.1.12 BoilerNet: Web Content Extraction 18

2.1.13 Brexit Analyzer: Vote Intent 18

2.1.14 Brexit Analyzer: Party Allegiance 20

2.1.15 GATE Hate: Abuse toward UK MPs 21

2.2 Shared Corpora **21**

2.2.1 NERD Tweet Corpus 21

2.2.2 Rumours Dataset 22

3 Unsupervised Evaluation **24**

3.1 Use Cases **24**

3.1.1 Trip Builder 24
3.1.2 MaxAndSam Network Reconstruction Method 25
3.1.3 DebtRank Systemic Risk Estimation Method 26
3.1.4 Generalized Network Dismantling 26
3.1.5 Maximum-Entropy network reconstruction 27
3.1.6 Network construction via tail Granger-causality 28
3.1.7 DEMON 28
3.1.8 TILES 28
3.2 Software 29
3.2.1 Egonetworks 29
4 Simulation and Synthetic Data 31
4.1 Use Cases 31
4.1.1 Carpooling - Carpooling Network Analysis 31
4.1.2 Soccer teams ranking simulator 31
4.2 Methods/Software 32
4.2.1 Ditras: Diary-based Trajectory Generator 32
4.2.2 NDlib/NDlib-REST 32
5 Shared Tasks and Evaluation Frameworks 33
5.1 Participation 33
5.1.1 TagMe and WAT: Entity discovery in texts 33
5.1.2 Hyperpartisan News 33
5.1.3 SMAPH: Entity discovery in queries 34
5.2 Competitions Organized 35
5.2.1 RumourEval 35
6 Conclusions 37
REFERENCES 38

DELIVERABLE SUMMARY

D11.3 is the second deliverable focusing on the SoBigData evaluation framework toolkit and datasets. The deliverables present the SoBigData evaluation data collection toolkit, which enables campaign participants access to the evaluation datasets, as described in T11.2. In addition, the deliverables comprise 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).

D11.3 builds on D11.2 by reporting additional work in a manner that facilitates task definition and comparison of algorithms and approaches, and presenting evaluation frameworks and datasets, as discussed in T11.2 (access to datasets is provided by the SoBigData Gateway).

EXECUTIVE SUMMARY

SoBigData's core mission is to create a 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". As an open research infrastructure, SoBigData promotes repeatable and open science. The work presented here supports this goal through two key foci: infrastructure for the effective and responsible sharing of data (the "framework"), and the resources and experiences that have arisen within the five thematic areas of the project: social media mining; social network analysis; human mobility analytics; web analytics; and visual analytics. Work here is presented under four headings pertaining to diverse task definitions that have emerged through the project; supervised evaluation, unsupervised evaluation, simulation and synthetic data, and shared tasks and evaluation frameworks. In addition to describing their contribution to task definition, partners share resources that can be accessed through the SoBigData Gateway subject to appropriate conditions and licensing.

1 RELEVANCE TO SOBIGDATA

As an open research infrastructure, SoBigData promotes repeatable and open science. Advancing the state of the art with regards to evaluation methods and infrastructure, and developing high utility shared resources for evaluation is a crucial part of this. Various work in the area of evaluation has been taking place under the umbrella of SoBigData. This deliverable presents this in the form of a taxonomy.

1.1 PURPOSE OF THIS DOCUMENT

We use the wide range of evaluations taking place under the SoBigData umbrella to taxonomize evaluation methods and reflect on them in practical terms. 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 containing the method name, objectives, and its key performance indicators. We also report the thematic cluster that it belongs to: Text and Social Media Mining (TSMM), Social Network Analysis (SNA), Human Mobility Analytics (HMA), Web Analytics (WA), Visual Analytics (VA) or Social Data (SD).

1.2 RELEVANCE TO PROJECT OBJECTIVES

SoBigData opens 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. Effective community evaluation efforts are a crucial part of this.

1.3 SOBIGDATA PROJECT DESCRIPTION

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”. 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. It will not only strengthen the existing clusters of excellence in social data mining research, but also create a pan-European, inter-disciplinary community of social data scientists, fostered by extensive training, networking, and innovation activities.

In addition, as an open research infrastructure, SoBigData promotes repeatable and open science. Although SoBigData is primarily aimed at serving the needs of researchers, the openly available datasets and open source methods and services provided by the new research infrastructure will also impact industrial and other stakeholders (e.g. government bodies, non-profit organisations, funders, policy makers).

1.4 RELATION TO OTHER WORKPACKAGES

The work package builds on D11.2 to bring together the information in the form of an analysis of evaluation types, and supplement that with our work in the area of promoting effective evaluation, namely by also

including shared corpora, tasks that have been organised under the umbrella of SoBigData and our work developing and participating in evaluation frameworks.

1.5 STRUCTURE OF THE DOCUMENT

In the next sections we survey the evaluation methods and resources in the SoBigData platform grouping them by evaluation approach. First we consider supervised evaluations, the way these have been approached within SoBigData and the corpora that have been shared. Then we consider the work that has been done in the area of unsupervised evaluation. Then we cover simulations and synthetic data. Finally we present work in the area of shared tasks and evaluation frameworks.

2 SUPERVISED EVALUATION

In this section we describe work in which supervised evaluation methods were employed; namely, method output was compared to manually annotated gold standard data.

2.1 USE CASES

2.1.1 MYWAY - TRAJECTORY PREDICTION

Exploratory: City of Citizens
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 behaviours modelled 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 strategy is a combination of the previous two. MyWay only requires sharing 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 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].

2.1.2 HUMAN MOBILITY DATA PRIVACY RISK ESTIMATOR

Exploratory: City of Citizens
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]. We construct a classification training dataset TC for every distinct background knowledge (this means that, in our experiments, we build a total of 33 distinct classification training datasets). Every classification dataset TC is used to train a classifier M using Random Forest. We evaluate the overall performance of a classifier by two metrics: (i) the accuracy of classification (ACC), and (ii) the weighted average F-measure. All the experiments are performed using a k-fold cross validation procedure with k=10. Here, we do not report the whole experimental results, but we only highlight that both evaluation metrics reach good results. For example, the maximum performance values reached by a

classifier are ACC= 0.95 and F-measure = 0.95, while the lowest performance values are ACC = 0.62 and F-measure = 0.59. However, in both cases we have significant improvements w.r.t. baseline.

2.1.3 STATISTICALLY VALIDATED NETWORKS

Exploratory: City of Citizens
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.

2.1.4 MACHINE LEARNING PREDICTION OF VOLATILITY IN BITCOIN

Exploratory: Well-being & Economy
Thematic Cluster: HMA, SNA
Partners Acronym: ETHZ
Dataset Used: Bitcoin market data

Evaluation: This is a machine-learning model that predicts the level of price fluctuations in the next hour segment. In particular, we study the problem of the Bitcoin short-term volatility forecasting based on volatility history and order book data. Order book, consisting of buy and sell orders over time, reflects the intention of the market and is closely related to the evolution of volatility. We propose temporal mixture models capable of adaptively exploiting both volatility history and order book features. By leveraging rolling and incremental learning and evaluation procedures, we demonstrate the prediction performance of our model as well as studying the robustness, in comparison to a variety of statistical and machine learning baselines. Meanwhile, our temporal mixture model enables to decipher the time-varying effect of order book features on volatility. It demonstrates the prospect of our temporal mixture model as an interpretable forecasting framework over heterogeneous Bitcoin data.

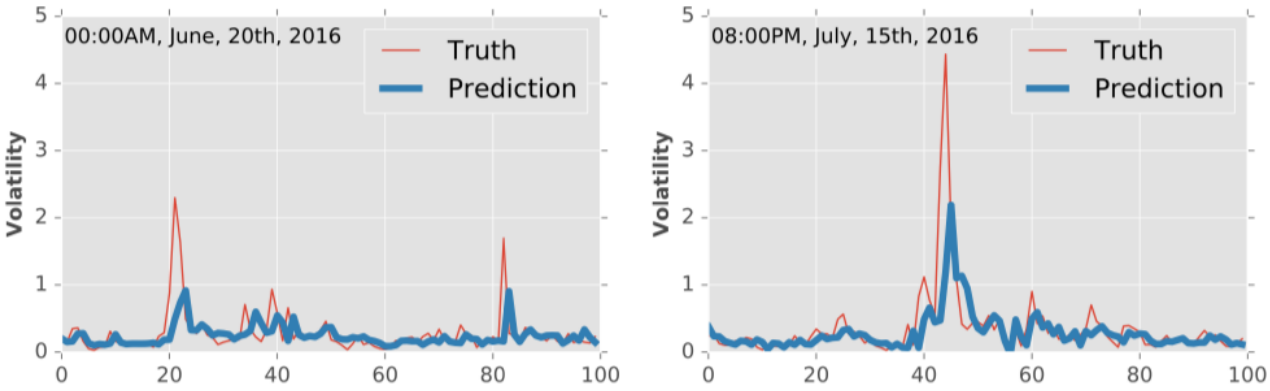


Figure 1 Prediction and true values of two sample periods.

2.1.5 MACHINE LEARNING PREDICTION OF VOLATILITY IN BITCOIN

Exploratory: Explainable Machine Learning

Thematic Cluster: HMA, SNA

Partners Acronym: ETHZ

Dataset Used: Meteorological data in Beijing of China & Energy production of a photo-voltaic power plant in Italy

Evaluation: For recurrent neural networks trained on time series with target and exogenous variables, in addition to accurate prediction, it is also desired to provide interpretable insights into the data. We explore the structure of LSTM recurrent neural networks to learn variable-wise hidden states, with the aim to capture different dynamics in multi-variable time series and distinguish the contribution of variables to the prediction. With these variable-wise hidden states, a mixture attention mechanism is proposed to model the generative process of the target. Then we develop associated training methods to jointly learn network parameters, variable and temporal importance w.r.t the prediction of the target variable. Extensive experiments on real datasets demonstrate enhanced prediction performance by capturing the dynamics of different variables. Meanwhile, we evaluate the interpretation results both qualitatively and quantitatively. It exhibits the prospect as an end-to-end framework for both forecasting and knowledge extraction over multi-variable data.

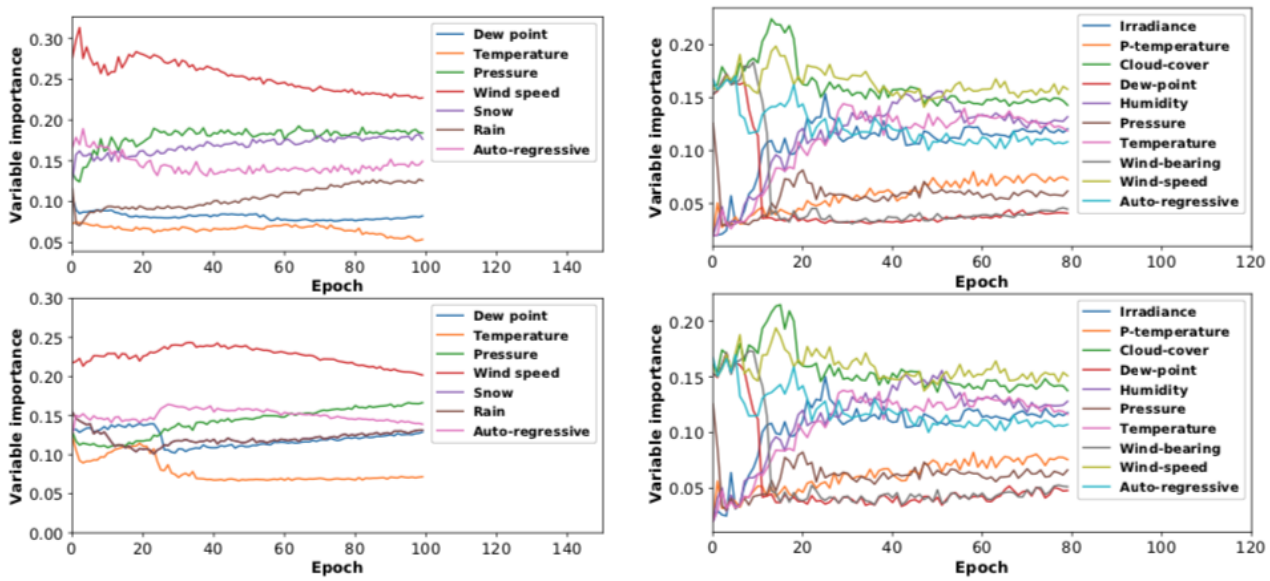


Figure 2 Variable importance over epochs during the training

2.1.6 SOCIOMETER

Exploratory: City of Citizens
Thematic Cluster(s): HMA
Partners Acronym: SoBigData.it - CNR
Datasets 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 person 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.

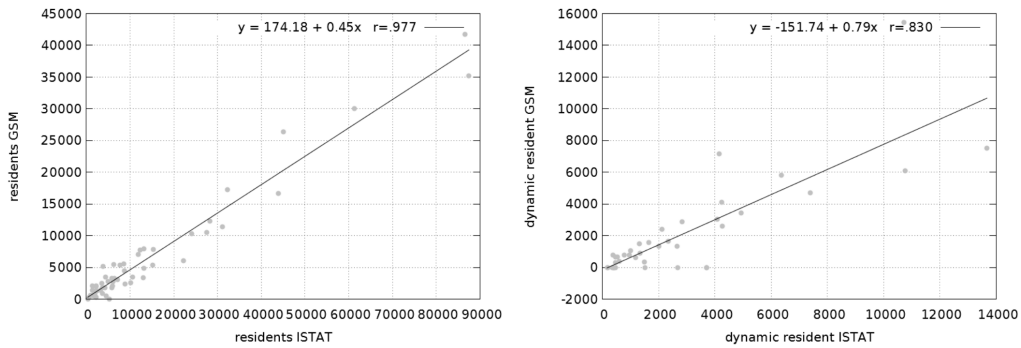


Figure 3 Comparison between Sociometer results and the official statistics

2.1.7 PRIVACY RISK ON SOCIOMETER

Exploratory: City of Citizens
Thematic Cluster(s): HMA
Partners Acronym: SoBigData.it - CNR
Datasets used: CDR Data - Tuscany

Evaluation: Given the methodology described in D11.1 for extracting profiles, we can analyse the privacy risks of the users. 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 (i.e., the information 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.

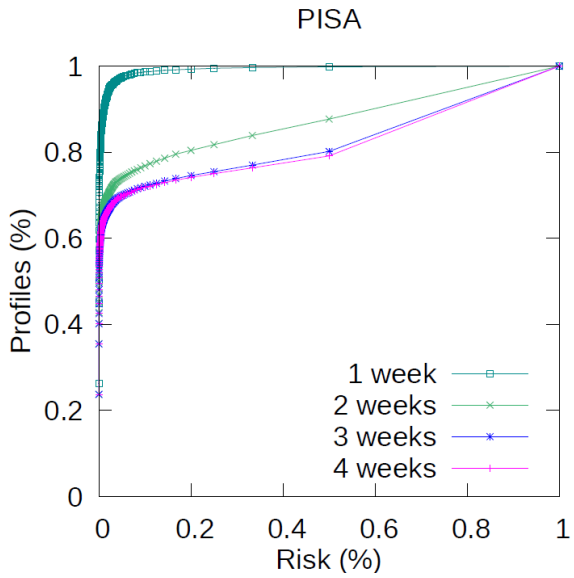


Figure 4 Cumulative distribution of Risk obtained with the attack simulation.

2.1.8 NEXT INSTITUTION PREDICTION BASED ON SCIENTIFIC PROFILE

Exploratory: Migration Studies
Thematic Cluster(s): HMA
Partners Acronym: SoBigData.it - CNR
Datasets used: Scientific Publications Dataset

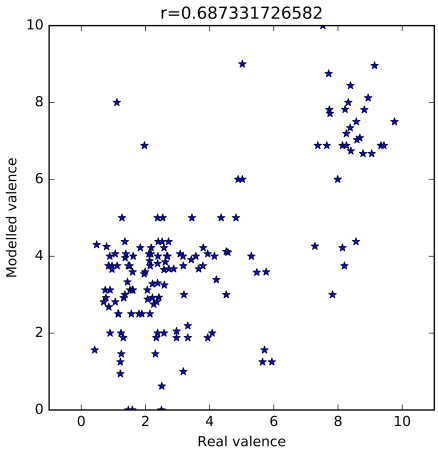
Evaluation: This method aims at predicting the future institution of a scientist given her recent scientific profile [36]. In the first phase, a data mining approach is used to predict whether or not a scientist will migrate, using logistic regression and decision trees. The method is evaluated using cross validation and obtains 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.

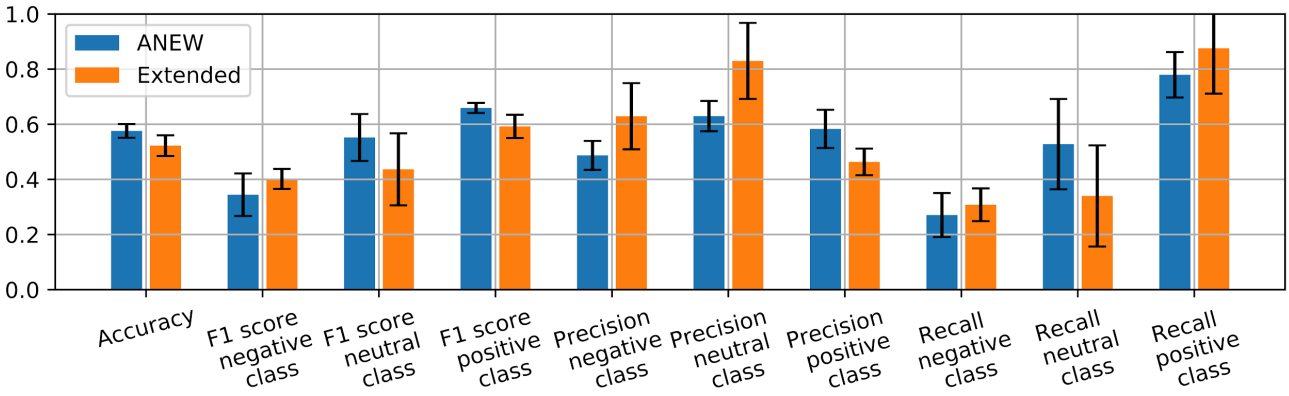
2.1.9 EPIDEMIC SENTIMENT ANALYSIS

Exploratory: Migration Studies
Thematic Cluster(s): TSMM
Partners Acronym: SoBigData.it - UNIPI, CNR.
Datasets 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 the correlation between the new dictionary and a manually annotated one (test dictionary). The resulting dictionary is shown to contain sentiment valences that correlate well with human-annotated sentiment, with values up to 0.7.



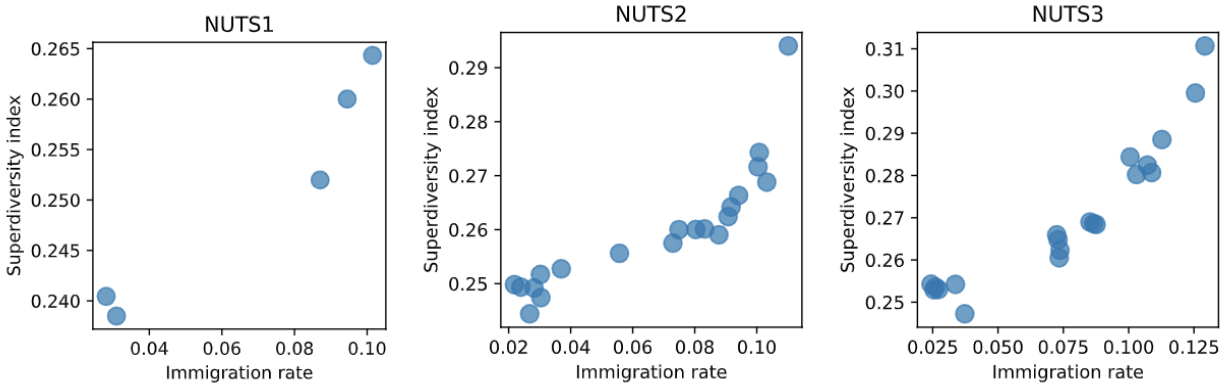
A further evaluation is based on classification of sentiment on Twitter, using the extended dictionary. We use the Semeval and Earth Hour datasets as gold standard, and we see results comparable to the original dictionary in terms of accuracy, recall, precision and F1-values. However we are able to tag more tweets compared to the original dictionary, which is an advantage of our method.



2.1.10 SUPERDIVERSITY AND SENTIMENT

Exploratory: Migration Studies
Thematic Cluster(s): TSMM
Partners Acronym: SoBigData.it - UNIPI, CNR.
Datasets used: Twitter Stream Dataset, D4I dataset

Evaluation: We have proposed a new index to measure superdiversity based on Twitter sentiment. This measures the sentiment valences of terms on twitter and compares them with an annotated dictionary. To evaluate the index, we compare it with high space resolution immigration rates from the D4I dataset. We show that our index correlates really well with immigrant stocks, in Italy and the UK, at various space resolutions. We obtain correlations between 0.8 and 0.95, results that are very promising in our quest to develop a nowcasting model of immigrant stocks. Below we show the relation between our superdiversity index and immigration rates in Italy at 3 different geographical levels.



2.1.11 SWAT: MYWAENTITY SALIENCE IN TEXTS

Exploratory: Cross-exploratory
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 thorough 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.

2.1.12 BOILERNET: WEB CONTENT EXTRACTION

Exploratory: Societal Debates
Thematic Cluster: WA
Partners Acronym: LUH
Dataset Used: CleanEval, GoogleTrends-2017 (self-created)

Evaluation: Web content extraction is an important task for numerous applications, ranging from usability aspects, like reader views for news articles in web browsers, to natural language processing or information retrieval. Existing approaches are tailored to a specific distribution of web pages, e.g. from a certain time frame, but lack in generalization power. We propose a neural model that takes only the HTML tags and words that appear in a web page as input. We show that our model matches the state-of-the-art performance on the CleanEval dataset. In addition, we create a new, more current dataset to show that our model is able to adapt to changes in the structure of web pages and outperform the state-of-the-art model.

2.1.13 BREXIT ANALYZER: VOTE INTENT

Exploratory: Societal Debates
Thematic Cluster: TSMM.
Partners Acronym: USFD
Dataset Used: Brexit Twitter User Vote Intent

Evaluation: The UK EU membership referendum created a focal point of international interest, and provides an opportunity to study dynamics around how belief propagation affects opinion, and how beliefs are introduced. Accurate classification of users into those who supported the UK leaving the EU and those that wished the UK to remain in the EU has supported key research objectives. The work has also provided a

sound basis for our input to the UK Fake News parliamentary subcommittee.[1] We were able to obtain accurate lists of “leavers” and “remainers” using an entirely automated approach, described in the next section. Classification of users according to referendum vote intent 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. Table 1 gives detailed statistics for three conditions; one matching tweet found for that user, two found or three found. “Total” is the total number of users found with that number of matching tweets. “Brndstr found” is the number of those users found in the Brndstr set, and so able to be evaluated.

	Total	**Brndstr found**	**Of found, correct**	**Accuracy**	**Cohen's Kappa**
Leavers, 3#	3539	1,142	1,129	0.987	0.972
Remainers, 3#	26,674	603	594		
Leavers, 2#	49,080	1,368	1,350	0.984	0.966
Remainers, 2#	50,972	901	882		
Leavers 1#	114,519	1,935	1,801	0.943	0.885
Remainers, 1#	175,042	1,744	1,667		

Table 1 *Brexit Classifier Accuracy*

[1] http://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/inquiries/parliament-2017/fake-news-17-19/

2.1.14 BREXIT ANALYZER: PARTY ALLEGIANCE

Exploratory: Societal Debates
Thematic Cluster: TSMM.
Partners Acronym: USFD
Dataset Used: UK General Election Vote Intent

Evaluation: A key piece of information in studying the recent UK general elections, as well as UK politics more broadly, is the accurate classification of users according to the political party they support. Hashtags were used to identify party supporters, as they were for Brexit vote intent. Negative hashtags weren't used, since opposing one party does not necessarily imply a vote intent for one particular other. Furthermore hashtags frequently used sarcastically are excluded. For example, Conservative slogan “strong and stable” was heavily used sarcastically. As for Brexit vote intent, tweets with such hashtags in the final position were used to identify party supporters, with thresholds of three, two and one such tweet being evaluated. Additionally, a further method considered party allegiance expressed in the Twitter biography.

A corpus was manually annotated to evaluate the work. On a sample of 51 bios annotated by three annotators, a Fleiss' Kappa interannotator agreement of 0.991 was achieved. 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. Accuracies of the automatic methods on this corpus are given in table 2.

Party	**Bios**	**1#+**	**2#+**	**3#+**
Labour Party	0.957	0.977	0.970	0.962
Conservative Party	0.798	0.923	1	1
Liberal Democrats	0.915	1	1	1
Scottish National Party	0.941	1	1	1
Plaid Cymru	No data	1	1	No data
Green Party	0.978	1	1	1
UKIP	0.952	0.978	0.957	1
Sinn Fein	1	No data	No data	No data

Democratic Unionist Party	1	No data	No data	No data

Table 2 *Party Allegiance Classifier Accuracy*

2.1.15 GATE HATE: ABUSE TOWARD UK MPS

Exploratory: Societal Debates
Thematic Cluster: TSMM.
Partners Acronym: USFD
Dataset Used: Kaggle "Detecting Insults in Social Commentary"

Evaluation: GATE's project detects verbal abuse in social media, and has been used extensively under SoBigData Societal Debates to investigate dialogue with political figures. The task is growing in importance, so a number of corpora are available and have been considered for evaluating our work. Jigsaw shared a corpus for their shared task (https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge). The corpus is annotated for identity characteristics, such as race and religion, allowing systems to be evaluated for presence of bias; that is, finding more false positives if the message recipient is for example black. However this corpus has a broader definition of abuse than most would consider practical. A similar issue was found with the OffensEval corpus. Precise definitions of abuse don't matter where systems are trained and tested on splits of the same corpus to demonstrate technological superiority. When it comes to evaluating a system for its success in a real world task however, it matters.

Kaggle's older corpus (https://www.kaggle.com/c/detecting-insults-in-social-commentary/data) has a more intuitive understanding of what constitutes abuse, and our system was improved by development against the training data. We achieve an accuracy of 0.81 and a precision/recall/F1 of 0.72, 0.47, 0.57 on the test data. Recent work by Wiegand et al (2019) offers a guide as to the state of the art for abuse detection systems tested *across domains*, to avoid the problem mentioned above, and finds a median F1 for well-known systems of 0.62. We prefer our rule-based approach as we avoid unintentional bias by not using indiscriminate features, we are able to easily add new words that become relevant, and the system is fast and stable on very large data. The approach is described in Gorrell et al (2019).

2.2 SHARED CORPORA

2.2.1 NERD TWEET CORPUS

Exploratory: Societal Debates
Thematic Cluster: TSMM
Partners Acronym: USFD

Dataset: The dataset comprises a set of 794 tweets annotated with named entities disambiguated against DBpedia. 400 of those were tweets from 2013 coming from financial institutions and news outlets, which were chosen due to the relatively high frequency of named entities within. They are challenging for entity

recognition and disambiguation, since capitalisation is not informative (all words have initial capital), but on the other hand, they are quite grammatical. The rest are random tweets collected in 2014, as part of the DecarboNet project on analysing online climate change debates. Keywords such as “climate change”, “earth hour”, “energy”, and “fracking” were used and the 394 tweets were chosen as a representative sample, containing sufficient named entities, without significant repetition. The tweets were annotated manually by a team of 10 NLP researchers, using a CrowdFlower interface. Each tweet was tagged by three annotators. Annotations for which no clear decision was made were adjudicated by a fourth expert, who had not previously seen the tweets. Unanimous inter-annotator agreement occurred for 89 % of entities, which can be used as the upper bound on performance attainable by an automatic method on this dataset and task. The resulting corpus contains 252 person annotations, 309 location annotations, 347 organization annotations and 218 nil annotations. The corpus and its creation are fully described in Gorrell et al (2015) and the corpus is available in the SoBigData portal.

2.2.2 RUMOURS DATASET

Exploratory: Societal Debates
Thematic Cluster: TSMM
Partners Acronym: USFD

Dataset: Automatic rumour verification is drawing research and commercial interest, as misleading rumours on the internet are increasingly associated with negative social outcomes. Our rumour verification corpus contains tweet threads in which the source tweet originates a rumour that may be true, false or unverified. Each thread then contains discussion in which the veracity of the rumour may be discussed. Figure 5 illustrates the structure of the corpus, using the root topic of the disappearance of Vladimir Putin as an illustration. Within that topic, there are several source tweets in which a rumour is originated, which may be true, false or unverified. (Examples given in the figure are that he is dead, he has the flu etc. Conceivably, the same rumour could be originated in several source tweets.) Below the source tweets is a discussion containing a potentially large number of tweets, each of which may support, deny, query or comment on the rumour. This discussion has been found helpful in determining veracity. The dataset originated as part of the PHEME project, and was extended under SoBigData to include new Twitter rumours (the test data) and Reddit rumours (new training and test corpora). The resulting enlarged corpus was the basis of the RumourEval 2019 task, discussed below under shared tasks. Full documentation can be found in Gorrell et al (2019). It contains a total of 446 rumour-originating tweets (across a smaller number of actual rumours) and 8574 responses, split into training and test sets.

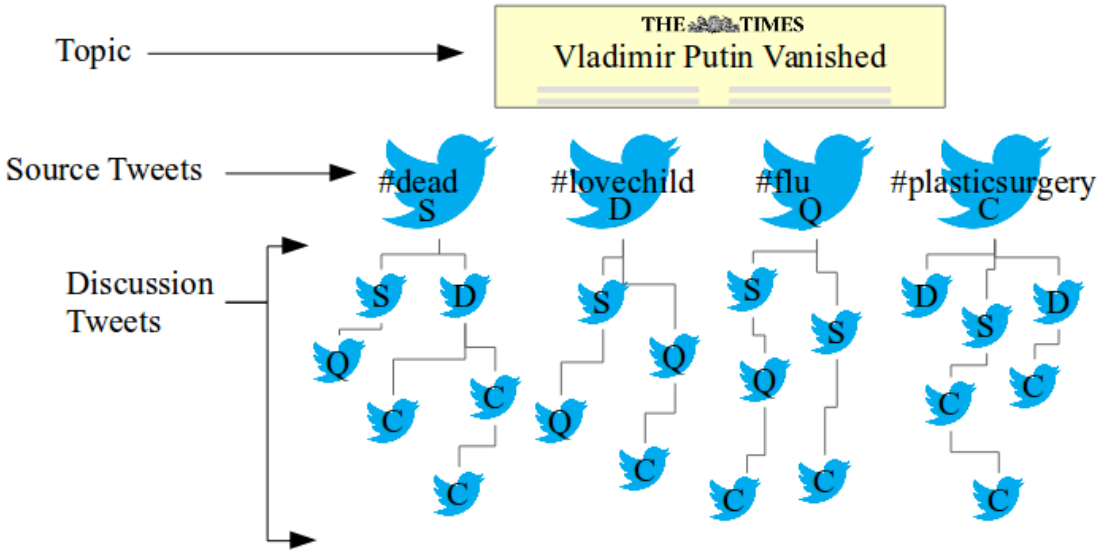


Figure 5 The structure of the rumours corpus

As well as forming the basis for two RumourEval shared tasks, the corpus is now the foundation of GATE's Twitter Rumour Verifier, which is available at **https://tweetveracity.gate.ac.uk**, and was presented at EMNLP 2019 (Karmakharm et al 2019).

3 UNSUPERVISED EVALUATION

In this section, work in which unsupervised methods were employed is described. Unsupervised methods may be evaluated heuristically, in terms of their utility in achieving the goals of the work.

3.1 USE CASES

3.1.1 TRIP BUILDER

Exploratory: City of Citizens
Thematic Cluster: HMA
Partners Acronym: SoBigData.it - CNR, HPC
Dataset Used: Flickr and Wikipedia Tourism Trajectories

Evaluation: TripBuiler 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 like so: (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 6. 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 6 Metrics for TripBuilder

3.1.2 MAXANDSAM NETWORK RECONSTRUCTION METHOD

Exploratory: Well-being & Economy
Thematic Cluster(s): SNA
Partners Acronym: SoBigData.it - IMT
Datasets 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 [5,6,55,56]. 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).

3.1.3 DEBTRANK SYSTEMIC RISK ESTIMATION METHOD

Exploratory: Well-being & Economy
Thematic Cluster(s): SNA
Partners Acronym: SoBigData.it - IMT
Datasets 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 also been tested in combination with the MaxAndSam reconstruction method, being accurately reproduced [3].

3.1.4 GENERALIZED NETWORK DISMANTLING

Exploratory: Well-being & Economy
Thematic Cluster(s): SNA
Partners Acronym: ETHZ
Datasets used: Social networks, Airport network

Evaluation: The proper functioning of many sociotechnical systems depends on their level of connectivity. By removing or deactivating a specific set of nodes, a network structure can be dismantled into isolated subcomponents, thereby disrupting the malfunctioning of a system or containing the spread of misinformation or an epidemic. We propose a generalized network-dismantling framework, which can take realistic removal costs into account such as the node price, the protection level, or removal energy. We discuss applications of cost-efficient dismantling strategies to real-world problems such as containing an epidemic or dismantling criminal or corruption networks.

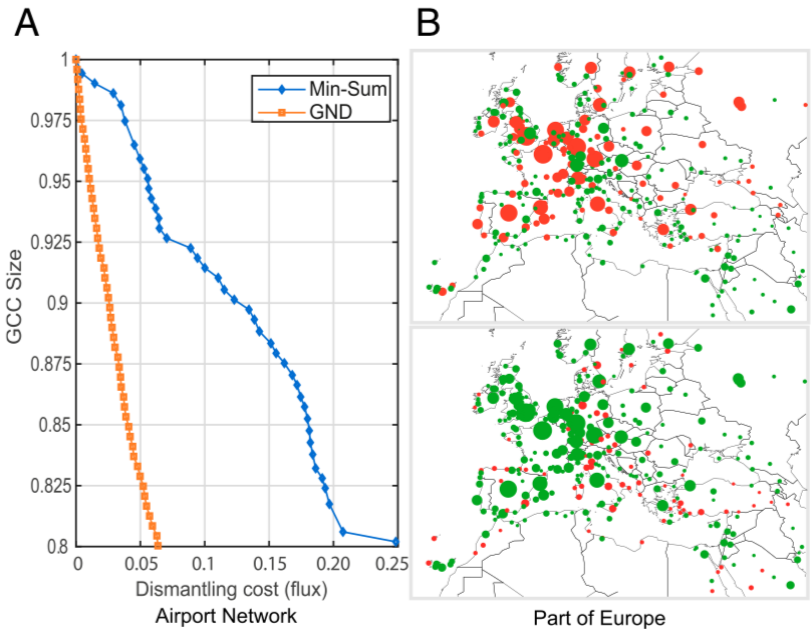


Figure 7 Comparison of the dismantling performance for the airport network, where the removal cost is the total passenger flow of the airport. (*A*) Setting the target size to *c* = 80%, the Min-Sum algorithm implies a cost of closing airports with ~25% of the total passenger flow. In contrast, our GND strategy dismantles the network to *c* = 80% size by a cost of only 6%. In *B*, red circles visualize the airports in Europe that were closed by the Min-Sum (*Upper*) or the GND (*Lower*).

3.1.5 MAXIMUM-ENTROPY NETWORK RECONSTRUCTION

Exploratory: Well-being & Economy
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.

3.1.6 NETWORK CONSTRUCTION VIA TAIL GRANGER-CAUSALITY

Exploratory: Well-being & Economy
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.

3.1.7 DEMON

Exploratory: Cross-exploratory
Thematic Cluster(s): SNA
Partners Acronym: SoBigData.it - CNR, UNIPI, KDD
Datasets 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.

3.1.8 TILES

Exploratory: Cross-exploratory
Thematic Cluster(s): SNA
Partners Acronym: SoBigData.it - CNR, UNIPI, KDD
Datasets 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 evaluates the adherence an identified partitioning to the expected one).

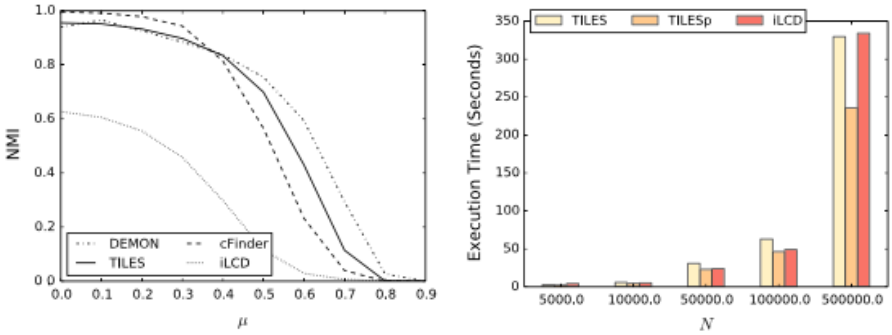


Figure 8 Dynamic Community Discovery: (left) NMI comparison, (right) TILES execution time.

Our results, highlighted in Fig. 8, 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)).

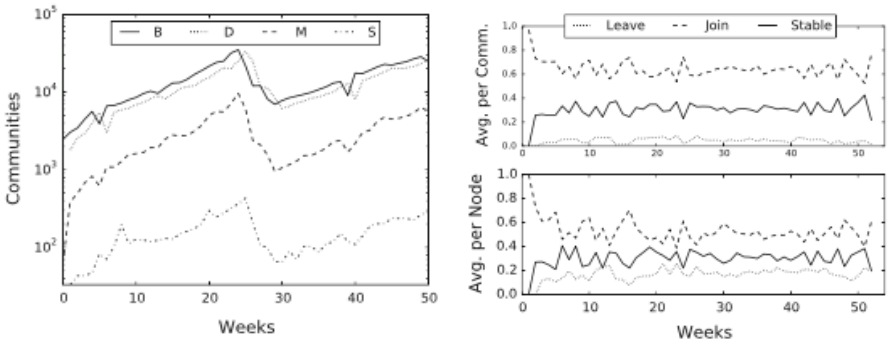


Figure 9 TILES Community Stability: (left) trends for community events - Birth, Merge, Split, Death; (right) example of trends for node/community stability.

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.

3.2 SOFTWARE

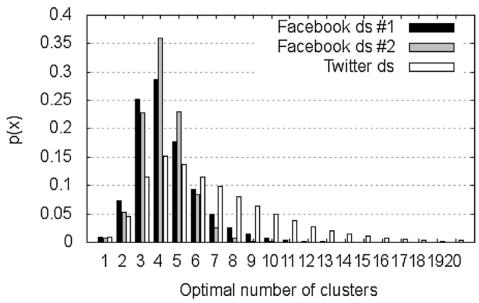
3.2.1 EGONETWORKS

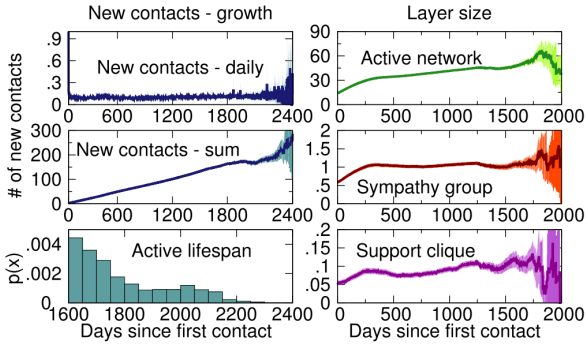
Exploratory: Cross-exploratory
Thematic Cluster: SNA
Partners Acronym: CNR

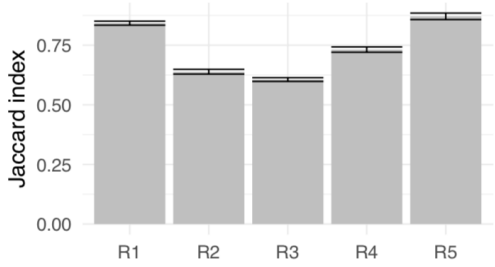
Datasets used: Facebook EuroSys 2009, Facebook - New Orleans regional network

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 (see figures below for an example).

The package has been used in the following recent publications: [50-53].







4 SIMULATION AND SYNTHETIC DATA

In this section, work is described in which it has been found appropriate to evaluate method output against synthetic data, or using simulation, in addition to work that supports this.

4.1 USE CASES

4.1.1 CARPOOLING - CARPOOLING NETWORK ANALYSIS

Exploratory: City of Citizens
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].

4.1.2 SOCCER TEAMS RANKING SIMULATOR

Exploratory: Sports Data Science
Thematic Cluster(s): HMA
Partners Acronym: SoBigData.it - Unipi
Datasets 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].

4.2 METHODS/SOFTWARE

4.2.1 DITRAS: DIARY-BASED TRAJECTORY GENERATOR

Exploratory: City of Citizens
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 development 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].

4.2.2 NDLIB/NDLIB-REST

Exploratory: Cross-exploratory
Thematic Cluster(s): SNA
Partners Acronym: SoBigData.it - CNR, UNIPI, KDD
Datasets used (Same name in catalogue): Synthetic graph generators

Evaluation: NDlib is a python library that allows to easily describe network diffusion simulations [17, 18]. NDlib allows one 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.

5 SHARED TASKS AND EVALUATION FRAMEWORKS

In this section we focus particularly on shared tasks and frameworks, as a way of improving evaluation by making it possible to compare methods more accurately.

5.1 PARTICIPATION

5.1.1 TAGME AND WAT: ENTITY DISCOVERY IN TEXTS

Exploratory: Societal Debates
Thematic Cluster: TSMM.
Partners Acronym: SoBigData.it – UNIPI
Dataset Used: GERBIL (http://aksw.org/Projects/GERBIL.html)

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 which a 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.

5.1.2 HYPERPARTISAN NEWS

Exploratory: Societal Debates
Thematic Cluster: TSMM
Partners Acronym: USFD
Dataset Used: Google Hyperpartisan News dataset

Evaluation: In 2019, Google organised a shared task under the auspices of SemEval, in which they shared a large corpus of news articles annotated for reliability of the source, and a smaller set of articles individually manually annotated for whether the article is highly partisan. The problem of the increasing proliferation of partisan news has been an ongoing focus for our work under SoBigData Societal Debates, so the task was considered highly relevant. Our system uses sentence representations from averaged word embeddings generated from the pre-trained ELMo model with Convolutional Neural Networks and Batch Normalization for predicting hyperpartisan news. The final predictions were generated from the averaged predictions of an ensemble of models. With this architecture, our system ranked in first place, based on accuracy, the official scoring metric. The work is described in Jiang et al, 2019 [49].

5.1.3 SMAPH: ENTITY DISCOVERY IN QUERIES

Exploratory: Cross-exploratory
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.

5.2 COMPETITIONS ORGANIZED

5.2.1 RUMOUREVAL

Exploratory: Societal Debates
Thematic Cluster: TSMM
Partners Acronym: USFD

Competition: Originally trading under the name SensEval, and focusing on word sense disambiguation, SemEval has expanded its remit to a range of semantic analysis tasks, and has run 11 times in the last 20 years, increasing from an original three-yearly cycle to running every year for the last six (though with the understanding that not every task will run every time). Usually, there are in excess of ten tasks each time, and there have been as many as 18, each attracting several or more participating teams. The workshops are co-located with major conferences, and publish formal proceedings, making them an attractive opportunity to showcase work. Including our evaluations under the SemEval umbrella is therefore an excellent opportunity to maximize impact. RumourEval is a shared task that ran in SemEval 2017 under the auspices of the PHEME project. We continued the series for SoBigData in 2019.

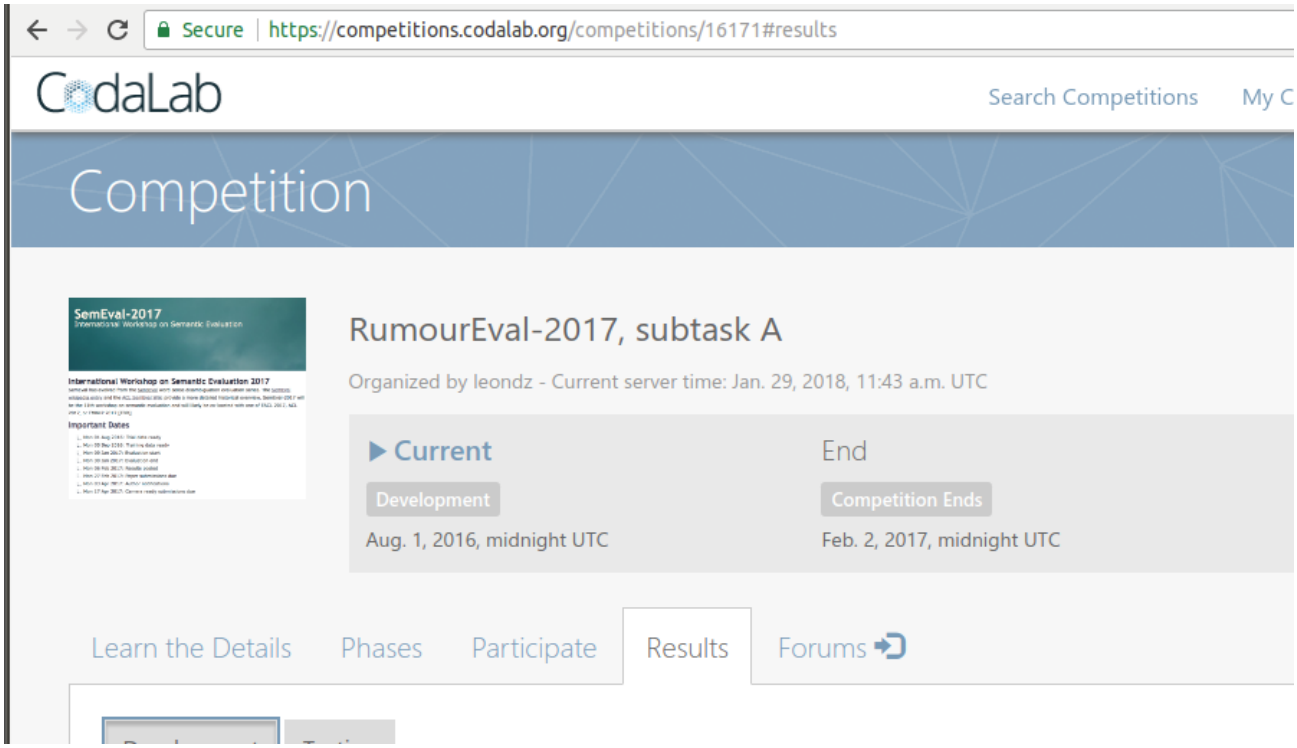


Figure 10 *CodaLab Competition Screenshot*

Figure 10 shows the CodaLab Competition page for the previous RumourEval task (described further below). The interface makes it easy to participate and view results. In figure 11, results against time are plotted automatically by CodaLab.

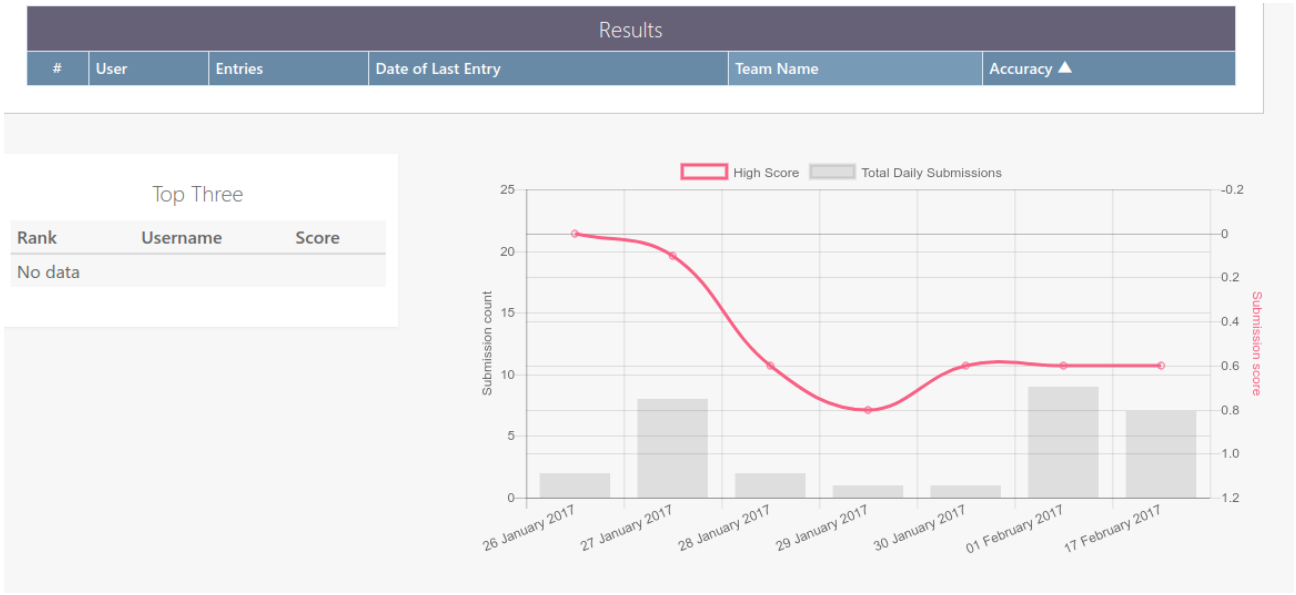


Figure 11 CodaLab Results Screenshot

6 CONCLUSIONS

Use cases, methods and other forms of contribution to evaluation have been explored here under four headings: supervised, unsupervised, simulation/synthetic and shared tasks/frameworks.

Under supervised evaluation, many partners have explored evaluation in the context of evaluating their research. MyWay aims to predict future data from past data, so the dataset was naturally leveraged to enable this. The Sociometer made use of official statistics. Epidemic Sentiment Analysis showed ingenuity in finding ways to evaluate the subjective matter of synonym quality. SWAT and GATE Hate operate in the area of established tasks, for which evaluation data already exists. The Brexit Analyzer leveraged available data in an ingenious way. A corpus was created to evaluate Party Allegiance. New corpora have also been shared by partners; a rumour verification corpus and a named entity recognition and disambiguation corpus.

Work requiring unsupervised evaluation mainly focused on networks. Trip Builder evaluates solutions found using heuristic metrics related to the utility of the solution in the real world. The MaxAndSam and DEMON methods are evaluated through practical comparison with other approaches. Whilst some supervised evaluation is possible, both DEMON and TILES make use of heuristic evaluation. An example of this is execution time. DebtRank shows practical utility in that it has been taken up by users. Egonetworks shares a library of software for network analysis.

In the area of simulation and synthetic data, several partners have demonstrated explorations through their use cases. Carpooling recommendations are evaluated through simulated utility. Simulation of soccer teams' ranking and comparing this to real world data provides a way of evaluating the model. Ditras models human mobility for downstream purposes; the work is evaluated against real world data. NDLib provides network simulation software for use in other projects.

Several partners have participated in shared tasks. TagMe and WAT have been evaluated within the GERBIL named entity linking framework. USFD participated in Google's hyperpartisan news detection shared task in SemEval 2019, coming in first place. SMAPH won the ERD 2014 short track challenge. USFD took a lead role in organising RumourEval 2019, a shared task forming part of SemEval, in which a record number of teams gathered around the goal of verifying rumours in social media data.

Evaluation is a crucial part of promoting repeatable and open science. Other core aspects of the SoBigData mission were also touched on in the context of evaluation. Within the context of the Sociometer and MyWay, user privacy was evaluated. The GATE Hate work discusses the bias of evaluation data. In the Bitcoin work, the interpretability of models was considered, an issue that increasingly has ethical implications.

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., Miccichè, 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.

[44] Gorrell, G., Petrak, J., & Bontcheva, K. (2015, May). Using@ Twitter conventions to improve# lod-based named entity disambiguation. In *European semantic web conference* (pp. 171-186). Springer, Cham.

[45] Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., & Derczynski, L. (2019). SemEval-2019 Task 7: RumourEval 2019: Determining Rumour Veracity and Support for Rumours. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 845-854). Association for Computational Linguistics.

[46] T. Karmakharm, N. Aletras and K. Bontcheva (2019). Journalist-in-the-Loop: Continuous Learning as a Service for Rumour Analysis. In EMNLP (demo).

[47] Gorrell, G., Bakir, M. E., Greenwood, M. A., Roberts, I., & Bontcheva, K. (2019). Race and Religion in Online Abuse towards UK Politicians. arXiv preprint arXiv:1910.00920.

[48] Wiegand, M., Ruppenhofer, J., & Kleinbauer, T. (2019, June). Detection of Abusive Language: the Problem of Biased Datasets. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 602-608).

[49] Jiang, Y., Petrak, J., Song, X., Bontcheva, K., & Maynard, D. (2019, June). Team Bertha von Suttner at SemEval-2019 Task 4: Hyperpartisan News Detection using ELMo Sentence

Representation Convolutional Network. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 840-844).

[50] Boldrini, C., Toprak, M., Conti, M., & Passarella, A. (2018, April). Twitter and the press: an ego-centred analysis. In Companion Proceedings of the The Web Conference 2018 (pp. 1471-1478). International World Wide Web Conferences Steering Committee.

[51] Arnaboldi, V., Passarella, A., Conti, M., & Dunbar, R. (2017). Structure of Ego-Alter Relationships of Politicians in Twitter. Journal of Computer-Mediated Communication, 22(5), 231–247.

[52] Arnaboldi, V., Dunbar, R. I. M., Passarella, A., & Conti, M. (2016). Analysis of Co-authorship Ego Networks. In LNCS - Advances in Network Science (pp. 82–96). Springer, Cham.

[53] 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.

[54] Pellungrini, Roberto, Pappalardo, Luca, Pratesi, Francesca, & Monreale, Anna. (2018, June). "Analyzing Privacy Risk in Human Mobility Data". In *Federation of International Conferences on Software Technologies: Applications and Foundations* (pp. 114-129). Springer, Cham.

[55] Amanah Ramadiah, Amanah, Caccioli, Fabio & Fricke, Daniel. (2018, September). “Reconstructing and stress testing credit networks”. ESRB Working Paper Series, No 84 / September 2018.

[56] Lebacher, Michael, Cook, Samantha, Klein, Nadja & Kauermann, Göran. (2019). “In Search of Lost Edges: A Case Study on Reconstructing Financial Networks”, arXiv:1909.01274