Issue 9 / Spring 2023

SOBIGDATA RESEARCH INFRASTRUCTURE

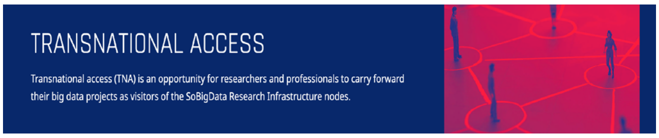
SOBIGDATA

RESEARCH INFRASTRUCTURE

Magazine

OPEN CALL

TransNational Access: a program of Short-Term Scientific Missions to carry forward your own big data project



TRANSNATIONAL ACCESS

Transnational access (TNA) is an opportunity for researchers and professionals to carry forward their big data projects as visitors of the SoBigData Research Infrastructure nodes.

[Continues on pag. 3]

Inside this issue

TNA Open Call

News

Challenge Us

Research Highlights

TransNational Access Highlights

Exploratories Highlights

TransNational Access: a program of Short-Term Scientific Missions to carry forward your own big data project

We welcome applications from individuals with a scientific interest, professionals, startups and innovators that may benefit from training in data science and social media analytics.

TRANSNATIONAL ACCESS

Transnational access (TNA) is an opportunity for researchers and professionals to carry forward their big data projects as visitors of the SoBigData Research Infrastructure nodes.

WHY APPLY

SoBigData RI manages vertical, thematic environments, called Exploratories, on the top of the SoBigData infrastructures for cross-disciplinary social mining research.

TNA provides researchers and professionals access to big data computing platforms, big social data resources, and cutting-edge computational methods of the selected Exploratory and enables multi-disciplinary social mining experiments with the SoBigData Research Infrastructure assets: big data sets, analytical tools, services, and skills.

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

WHO CAN APPLY

We welcome applications from individuals with a scientific interest, professionals, startups, and innovators who may benefit from data science and social media analytics training.

WHAT WE OFFER

Funding for a TNA participant is up to 5000 euros available to cover the cost of daily subsistence, accommodation, and

APPLY NOW!

Visit our website
http://www.sobigdata.eu/tna_call2023

economy flights/train.

WHEN APPLY

Transnational Access (TNA) activities and applications can be submitted anytime. Please note all applications received between July 15 and August 31 will not be processed until after September 1st.

Applications from female scientists are particularly encouraged.

For info about **HOW TO APPLY** please visit our website

http://www.sobigdata.eu/tna_call2023

Photo credit: Joshua Woroniecki - Pixabay

SoBigData Summer School

Responsible Data Science for Society: Models, Algorithms, Trustworthy AI

9-15 July 2023
Lipari island, Sicily -Italy
https://sobigdata23.liparischool.it

CALL FOR PARTICIPATION

ABSTRACT

Data Science and AI play an increasingly important role in our daily life. AI, with its applications, are essential tools for the ethical and responsible progress of our multicultural and interconnected society. The school for "Responsible Data Science for Society: Models, Algorithms, Trustworthy AI" introduces participants to selected topics to better understand the complexity of our world from the data scientist's perspective.

REGISTRATION

The registration fee is 600 €. The fee covers the course material, bus+hydrofoil Catania airport-Lipari-Catania airport, social events, and coffee breaks. Late registration is 700 €.
Early registration applications can be submitted up to May 31st, 2023. Late registrations will be accepted up to June 30th, 2023.

VENUE

The conference room (located at Hotel Giardino sul Mare, Via Maddalena, Lipari) is air-conditioned and equipped with all conference materials. In addition, special areas are reserved for students for the afternoon coursework and study. The island of Lipari can be easily reached from Milazzo, Palermo, Naples, Messina, and Reggio Calabria by ferry or hydrofoil (50 minutes from Milazzo).

GENERAL CHAIRS

Mark Cotè (KCL - UK) | Roberto Trasarti (ISTI-CNR)

ORGANIZING COMMITTEE

Marco Braghieri (KCL - UK) | Valerio Grossi (ISTI-CNR) | Michela Natilli (ISTI-CNR) | Beatrice Rapisarda (ISTI-CNR)

SPONSORS

This event is supported by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu) and - NextGenerationEU - National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Avviso n. 3264 del 28/12/2021

Responsible Data Science for Society: Models, Algorithms, Trustworthy AI

July 9th - July 15th, 2023

Summer School

Computational Complex and Social System: Spreading and Accessing Information

16-22 July 2023
Lipari island, Sicily -Italy
https://complex23.liparischool.it

CALL FOR PARTICIPATION

ABSTRACT

The edition of 2023 of the PhD Lipari School on Complex and Social Systems, held in Lipari, will deal with the topic "Spreading and accessing information" coming from many sources and with the use of several models of computations. The PhD school will see the participation of many prestigious speakers coming from all around the world. UniPI is co-organizer of the event since many years now, and includes some of the project participants as school co-directors and members of the advisory board.

TOPICS 2023

Models for epidemic and prevention measures, private retrieval of data, quantum security and teleportation, network algorithms to model biological information, probabilistic methods for microbiome dynamics, internet of things/people, smart cities.

REGISTRATION

The registration fee is 600 €. The fee covers the course material, bus+hydrofoil Catania airport-Lipari-Catania airport, social events, and coffee breaks. Late registration is 700 €.
Early registration applications can be submitted up to May 31st, 2023. Late registrations will be accepted up to June 30th, 2023.

VENUE

The conference room (located at Hotel Giardino sul Mare, Via Maddalena, Lipari) is air-conditioned and equipped with all conference materials. In addition, special areas are reserved for students for the afternoon coursework and study. The island of Lipari can be easily reached from Milazzo, Palermo, Naples, Messina, and Reggio Calabria by ferry or hydrofoil (50 minutes from Milazzo).

DIRECTORS

Alfredo Ferro (UNICT) | Paolo Ferragina (UNIPI) | Dirk Helbing (ETH, Zurich) | Carlo Ratti (MIT, USA)

SPONSORS

This event is supported by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu)

Spreading and Accessing Information

July 16th - July 22nd, 2023

Prompting and other algorithmic curiosities

Digital Methods Summer School and Data Sprint 2023

3-14 July 2023
Amsterdam, the Netherlands
https://wiki.digitalmethods.net/Dmi/SummerSchool2023

CALL FOR PARTICIPATION

The Digital Methods Initiative (DMI), Amsterdam, is holding its annual Summer School on 'Prompting and other algorithmic curiosities'. The format is that of a (social media and web) data sprint, with tutorials as well as hands-on work for telling stories with data. There is also a programme of keynote speakers. It is intended for advanced Master's students, PhD candidates and motivated scholars who would like to work on (and complete) a digital methods project in an intensive workshop setting.

REGISTRATION

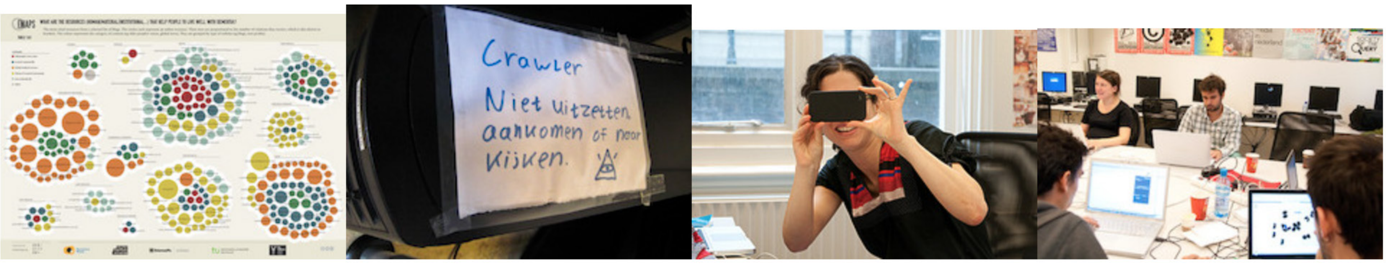
There are rolling admissions and applications are accepted until 15 May 2023. To apply please send a letter of motivation, your CV, a headshot photo, 100-word bio as well as a copy of your passport (details page only) to summerschool [at] digitalmethods.net. Notifications of acceptance are sent 1-2 weeks after application. Final notifications on 16 May. The full program and schedule of the Summer School are available by 20 June 2023.

The fee for the Digital Methods Summer School 2023 is EUR 895, and upon completion all participants receive transcripts and certificates (worth 6 ECTS). To complete the Summer School successfully all participants must co-present the weekly final presentations and co-author the weekly final project report, evidenced by the presentation slides or poster as well as the final report(s) themselves. Final reports should appear on this wiki (handy template) and contain a link to the final presentation slides or poster. They are due four weeks after the end of the Summer School. There are no other attendance or completion certificates issued other than the transcripts.

For further information, visit the website: https://wiki.digitalmethods.net/Dmi/SummerSchool2023

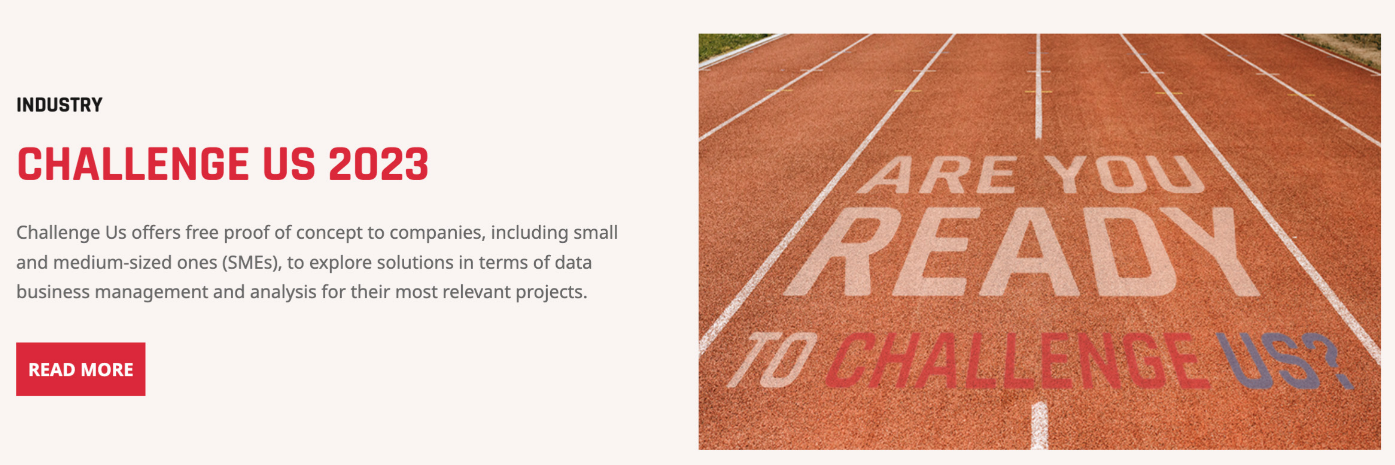
SPONSORS

This event is supported by the European Union – Horizon 2020 Program under the scheme "INFRAIA-01-2018-2019 – Integrating Activities for Advanced Communities", Grant Agreement n.871042, "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu)



Challenge Us 2023: SoBigData meets Industry

The Challenge Us program provide opportunities to companies interested in harvesting their own data they bring, thus entering the "world of Big Data" and exploit its potential. The Challenge Us is designed to build a bridge between industry and academia, offering the free support of SoBigData++ scientists to design solutions and produce proof of concepts.



INDUSTRY

CHALLENGE US 2023

Challenge Us offers free proof of concept to companies, including small and medium-sized ones (SMEs), to explore solutions in terms of data business management and analysis for their most relevant projects.

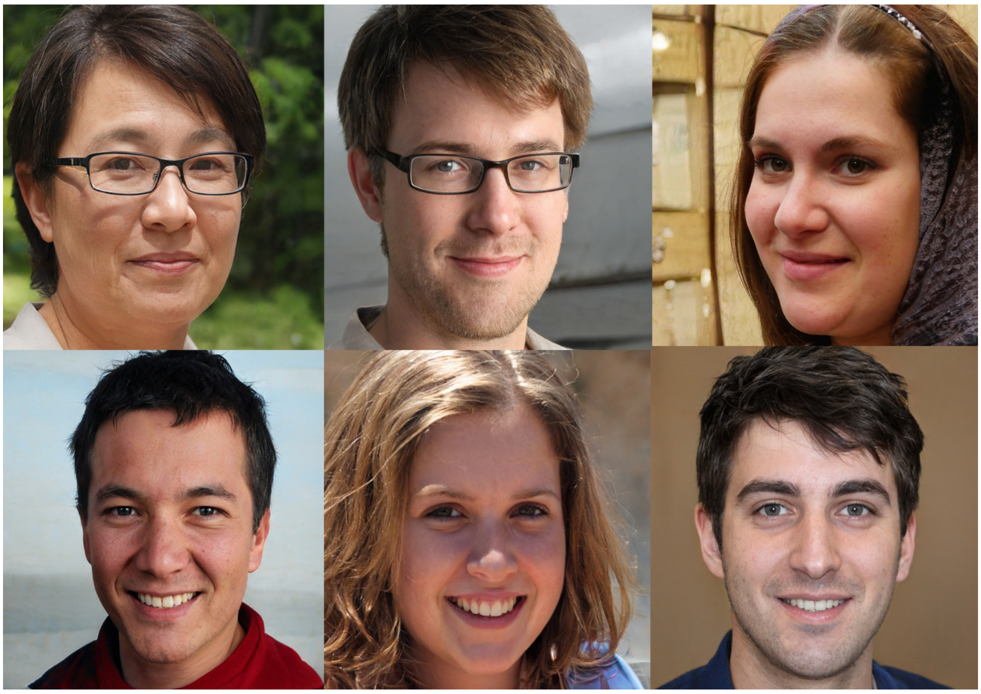
FOR MORE INFO:

http://www.sobigdata.eu/challenge-us-2023

Generating Synthetic Mobility Networks with Generative Adversarial Networks

This article explores the use of Generative Adversarial Networks to to study and generate daily mobility flows in a city, introducing MoGAN. As retrieving quality mobility data is hard, due the well-known privacy issue when dealing with such sensitive data, a tool such as MoGAN allows a useful way for performing high-level data-augmentation operations.

Giovanni Mauro, University of Pisa | g.mauro7@studenti.unipi.it



Look at these people. They look human, don't they? Well, too bad they do not exist, as these images are completely synthetic. This is possible thanks to a Deep Learning architecture called Generative Adversarial Networks (GANs) [R1].

These architectures are able to capture the probability distribution of a training set (of images, in this instance) and replicate it in order to create a new sample with the same probability distribution (therefore realistic) but not belonging to the training set. In a nutshell this architecture is made up of two building blocks. A Generator (G), an artificial neural network that takes as input a noisy vector and tries to fool a Discriminator (D, another neural network) by generating more and more realistic images. D has the task of distinguishing between synthetic images (generated by G) and realistic images (the one of the actual training set) by highly penalizing G if it is easily able to label the synthetic images as "fake" and by less penalizing G it is hard to distinguish synthetic and real images. You can find a scheme of a classic GAN in Figure 1.

What if we use this architecture to study and generate daily mobility flows in a city? Some definitions might be of use: first of all a tessellation is a partition of a city in (regular and squared in our case) zones called tiles. A Mobility Network (MN) is a weighted directed network in which nodes are tiles and edges represent the number of people moving between tiles. We represent a Mobility Network as a Weighted Adjacency Matrix (See Figure 2). But what have GANs to deal with Mobility Networks?Well, if you think about it a matrix can be seen as a mono-channel image (Figure 2), so here it's our intuition: *If we are able to generate synthetic images of different people, we can be able to generate synthetic matrices representing daily Mobility Networks of a city.*

We therefore introduce MoGAN (Mobility Generative Adversarial Network) [R3]. MoGAN is based on Deep Convolutional GAN (DCGAN) [R2] a particular type of GAN in which both G and D are Convolutional Neural Networks (CNNs): G performs an upsampling convolution (i.e., it takes a noisy vector as input and transforms it into a synthetic matrix), while D performs the classical convolutional classification. In our case, MoGAN will operate over a training set of daily Mobility Networks and at the end of the training process MoGAN's Generator will be able to generate as many fake Mobility Networks as desired (see Figure 3). DCGAN specifics require images (matrices) to be of dimension 64x64, this is why we split the city into 64 equally spaced squared tiles, so as to have the adjacency matrices of the mobility networks of dimension 64x64.

We trained MoGAN over 4 different datasets. MoGAN is now able to gen-

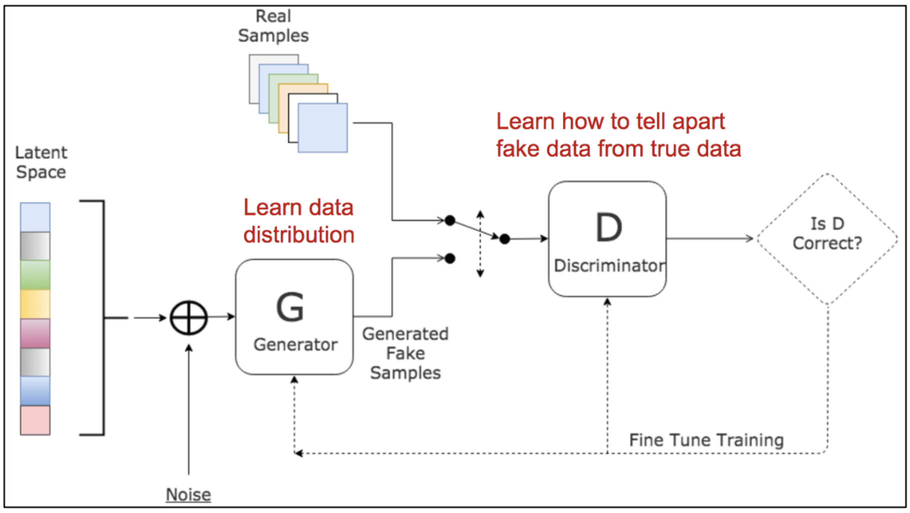
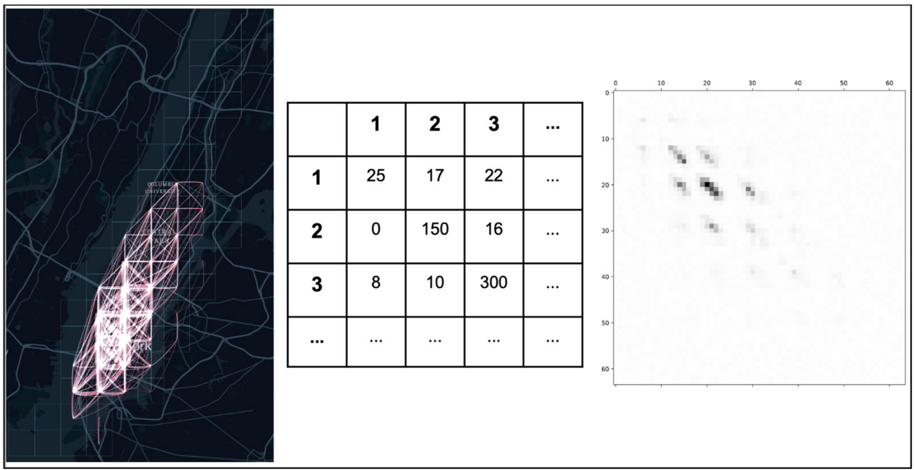


Figure 1: Classical GAN architecture.

erate synthetic networks of two different means of transport (bike and taxis) and of two different cities [New York City (NYC) and Chicago (CHI)]. For all of these 4 datasets, we transform the tabular data in input (each row containing information of the starting and ending zone of the bike or taxi ride) into a 64x64 adjacency matrix representation.

We do that by performing a spatial join operation with the tessellation of the city, grouping and counting rides starting and ending in each tile and transforming the dataset into a list of 64x64 daily arrays. A visual representation of the transformation into Mobility Networks for the NYC' bike dataset is given in Figure 4. Thus, MoGAN is now able to generate tons of realistic Mobility Networks. How do we evaluate its actual generative ability? In fact, evaluating if a face is realistic it's quite easy: we can look at the picture and decide if it is realistic or not. On the other hand, deciding if a network is similar to another is not so easy. In order to do so we compare our model with two classical mobility models for flow generation: Gravity[R4] and Radiation[R5] model. Gravity model postulate that the flows between two locations is inversely proportional to the distance between them, while the Radiation model considers the number of opportunities in each place, along with the distance, when generating flows.



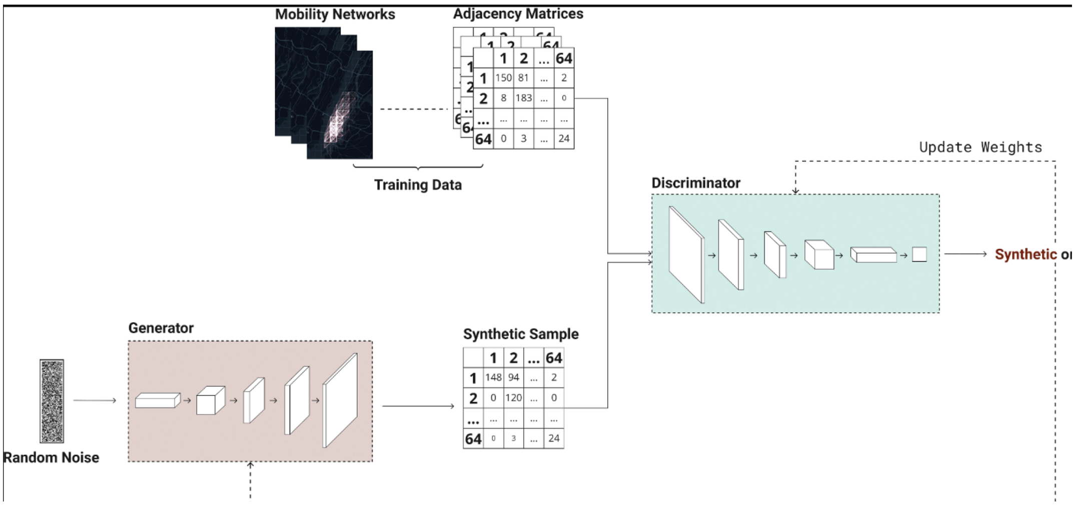
	1	2	3	...
1	25	17	22	...
2	0	150	16	...
3	8	10	300	...
...	...	...	...	...

Figure 2: From left to right: Visual, Matrix and b/w image representation of a daily MN.

For comparing our model with these baseline models, we create three sets:

- ● Test Set: A set of networks excluded from the training phase
- ● Synthetic Set: A set of fake networks generated by MoGAN
- ● Mixed Set: A set of networks coming half from the Training Set and half from the Synthetic Set.

After that, we calculate the distribution of several similarity measures (CPC, RMSE, JS divergence of the Weight distribution and several more). We would like the distribution over the three sets to be as much overlapping as possible. As you can see in Figure 5 that reports the results of the CPC analysis, the three distributions of our models are way more overlapping than the distributions of the other two models, for all of the 4 datasets.

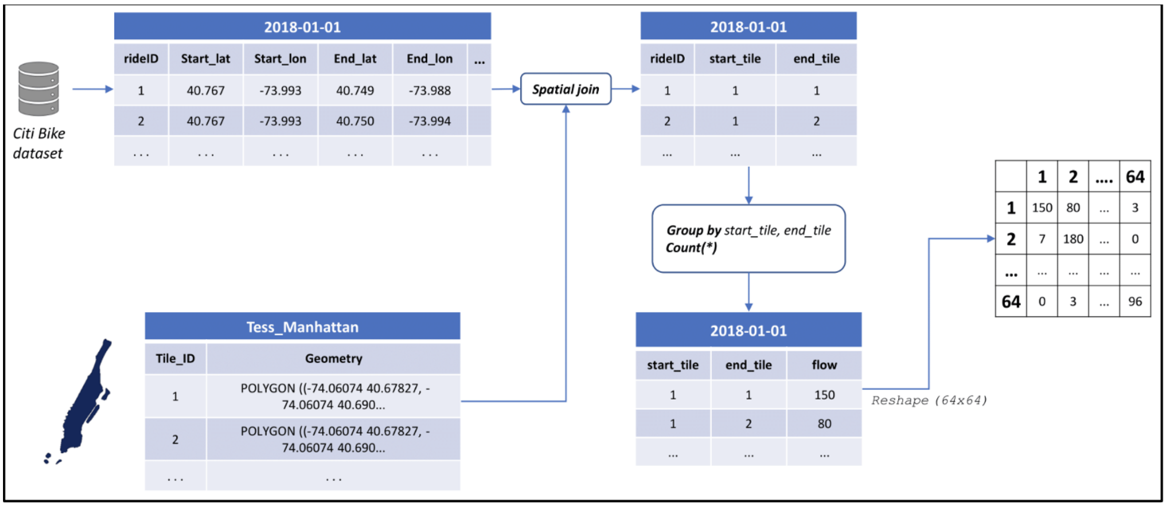


	1	2	...	64
1	150	81	...	2
2	8	183	...	0
...	...	...	...	...
64	0	3	...	24

	1	2	...	64
1	148	94	...	2
2	0	120	...	0
...	...	...	...	...
64	0	3	...	24

Figure 3: MoGAN architecture.

What's the scope of generating synthetic Mobility Networks? Well, there are



2018-01-01					
rideID	Start_lat	Start_lon	End_lat	End_lon	...
1	40.767	-73.993	40.749	-73.988	
2	40.767	-73.993	40.750	-73.994	
...	...	...	...	...	

2018-01-01		
rideID	start_tile	end_tile
1	1	1
2	1	2
...	...	...

Tess_Manhattan	
Tile_ID	Geometry
1	POLYGON ((-74.06074 40.67827, -74.06074 40.690...
2	POLYGON ((-74.06074 40.67827, -74.06074 40.690...
...	...

2018-01-01		
start_tile	end_tile	flow
1	1	150
1	2	80
...	...	...

	1	2	....	64
1	150	80	...	3
2	7	180	...	0
...	...	...	...	...
64	0	3	...	96

Figure 4: Data Extraction and Transforming phase for NYC' bike dataset.

several. Retrieving quality mobility data is hard, due the well-known privacy issue when dealing with such sensitive data. Therefore, a tool such as MoGAN allows a useful way for performing high-level data-augmentation operations. Furthermore, our model can be used as a useful what-if simulation tool.

P.S. Our model and analysis is completely reproducible, and can be found at https://github.com/jonpappalord/GAN-flow.

REFERENCES

[R1] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[R2] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." arXiv preprint arXiv:1511.06434 (2015).

[R3] Mauro, Giovanni, et al. "Generating Synthetic Mobility Networks with Generative Adversarial Networks." arXiv preprint arXiv:2202.11028 (2022).

[R4] Barbosa, Hugo, et al. "Human mobility: Models and applications." Physics Reports 734 (2018): 1-74.

[R5] Simini, Filippo, et al. "A universal model for mobility and migration patterns." Nature 484.7392 (2012): 96-100.

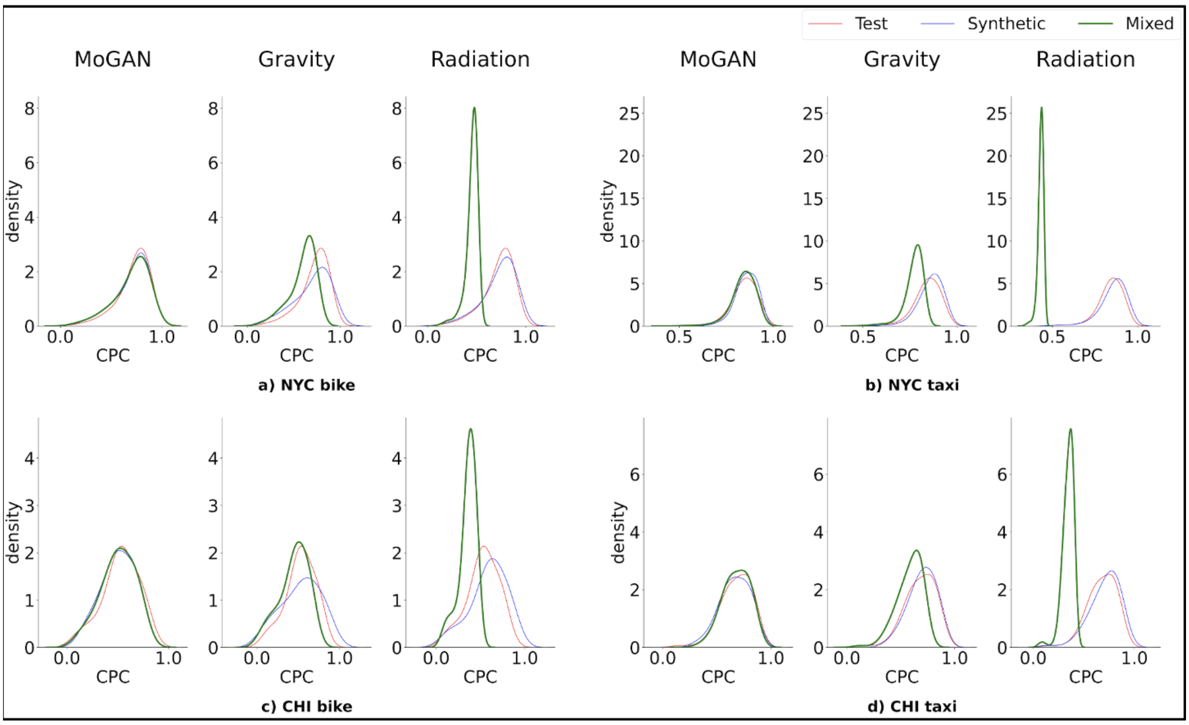


Figure 5: Distribution over the Test, Synthetic and Mixed set of the CPC scores of MoGAN over the three datasets

Blood sample profile helps to injury forecasting in elite soccer players

Alessio Rossi, University of Pisa

By analyzing external workloads with machine learning models (ML), it is now possible to predict injuries, but with a moderate accuracy. The increment of the prediction ability is nowadays mandatory to reduce the high number of false positives. The aim of this study was to investigate if players' blood sample profiles could increase the predictive ability of the models trained only on external training workloads.

Eighteen elite soccer players competing in Italian league (Serie B) during the seasons 2017/2018 and 2018/2019 took part in this study. Players' blood samples parameters (i.e., Hematocrit, Hemoglobin, number of red blood cells, ferritin, and sideremia) were recorded through the two soccer seasons to group them into two main groups using a non-supervised ML algorithm (k-means). Additionally to external workloads data recorded every training or match day using a GPS device (K-GPS 10 Hz, K-Sport International, Italy), this grouping was used as a predictor for injury risk. The goodness of ML models trained were tested to assess the influence of blood sample profile to injury prediction.

Hematocrit, Hemoglobin, number of red blood cells, testosterone, and ferritin were the most important features that allowed to profile players and to analyze the response to external workloads for each type of player profile. Players' blood samples' characteristics permitted to personalize the decision-making rules of the ML models based on external workloads reaching an accuracy of 63%. This approach increased the injury prediction ability of about 15% compared to models that take into consideration only training workloads' features. The influence of each external workload varied in accordance with the players' blood sample characteristics and the physiological demands of a specific period of the season.

Field experts should hence not only monitor the external workloads to assess the status of the players, but additional information derived from individuals' characteristics permits to have a more complete overview of the players well-being. In this way, coaches could better personalize the training program maximizing the training effect and minimizing the injury risk.

REFERENCES

https://link.springer.com/article/10.1007/s11332-022-00932-1

Photo courtesy by Alexander Fox | PlaNet Fox from Pixabay

Generation of complete realistic cellular network traffic

Anne Josiane Kouam, INRIA & Ecole Polytechnique, Palaiseau, France | anne-josiane.kouam-djuigne@inria.fr

Charging Data Records are acknowledged as a standard tool for studying human mobility, infrastructure usage, and traffic behavior. We name such datasets as CdRs to distinguish them from the traditional Call Detail Records (CDRs), describing call and SMS cellular communication only. CdRs describe time-stamped and geo-referenced event types (i.e., data, calls, SMS) generated by each mobile device interacting with operator networks. They comprise city-, region-, or country-wide areas and usually cover long periods (months or years); no other technology currently provides an equivalent per-device precise scope. As a result, CdRs represent a rich source of knowledge valuable to many communities such as sociology, epidemiology, or networking.

Yet, the exploitation of real-world CdRs for research faces many limitations. First, accessibility: CdRs datasets are not publicly available, imposing strict mobile operators' agreements. Second, usability: CdRs are usually available in an aggregated form (i.e., grouped mobility flows and coarse spatiotemporal information), limiting related analyses' preciseness. Third, privacy: even anonymized, non-aggregated CdRs describe sensitive information of users' habits, which hardens their shareability. This project aims to address such limitations by enabling the scientific community's autonomous generation of realistic and privacy-compliant CdRs, thus providing new avenues for research advances.

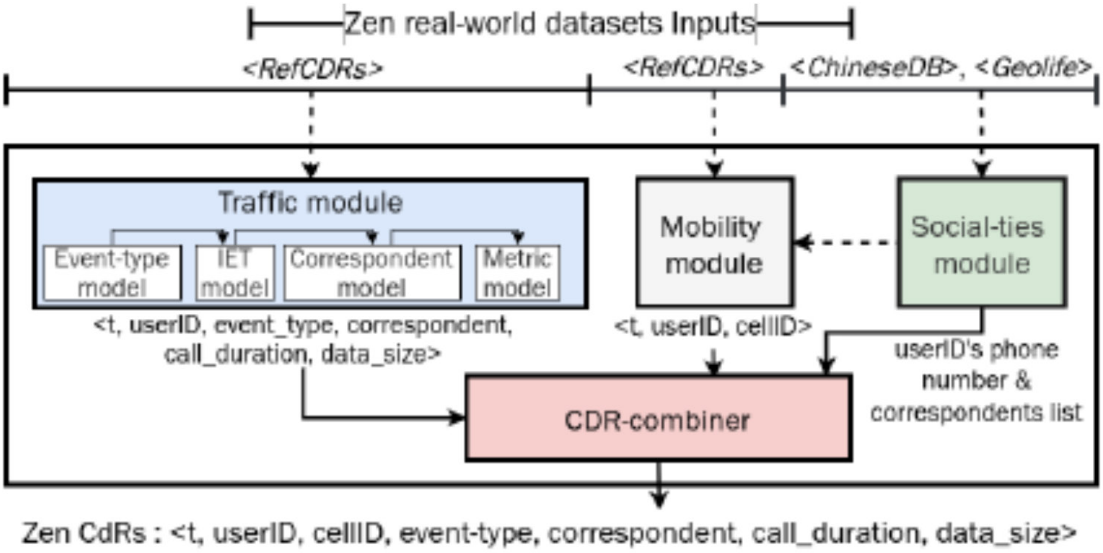

Zen real-world datasets Inputs
<RefCDRs>
<RefCDRs>
<ChineseDB>, <Geolife>
Traffic module
Event-type model
IET model
Correspondent model
Metric model
Mobility module
Social-ties module
<t, userID, event_type, correspondent, call_duration, data_size>
<t, userID, cellID>
userID's phone number & correspondents list
CDR-combiner
Zen CdRs : <t, userID, cellID, event-type, correspondent, call_duration, data_size>

Figure 1: Zen architecture

In particular, generated CdRs should conform to essential attributes, namely, completeness, realisticness, fine-grained description, and privacy, which makes the generation of realistic CdRs challenging and complex.

To respond to these criteria, we use as a baseline a previous framework (named Zen [R1]) with this same goal. An overview of the Zen framework is provided in Figure 2. Zen architecture consists of (1) a traffic module, (2) a mobility module, (3) a social-ties module, and (4) a CdR-combiner or merger module. The traffic module leverages Long-Short-Term Memory neural networks (LSTM) jointly with statistical analysis to model users' traffic behavior from real-world CdRs. The mobility module (i) emulates users' temporal displacements on a real-world geographical map over a selected period and (ii) associates corresponding users' positions with a real-world cellular topology. This dataset feeds the social-ties module that builds the network social structure on top of which users' communication interactions occur by creating users' phonebooks, i.e., a list of phone numbers a user is likely to contact. Finally, the CdR-combiner module combines all modules' outputs to generate realistic CdRs over a specified duration and particular urban area.

Despite the validated accuracy of Zen models to reproduce daily cellular behaviors of the urban population, Zen suffers some limitations related to the incompleteness of its reference datasets, whic have only traffic features and lack mobility ones. As a result, Zen fails to reproduce the distribution of the counts of generated events through time and unrealistically correlates users' traffic to mobility behaviors.

Our aim is to fix such Zen's limitations while extending it with the flexibility and generality to adapt its generation to target mobility zones with the help of complete real-world datasets from a major network operator in Chile.

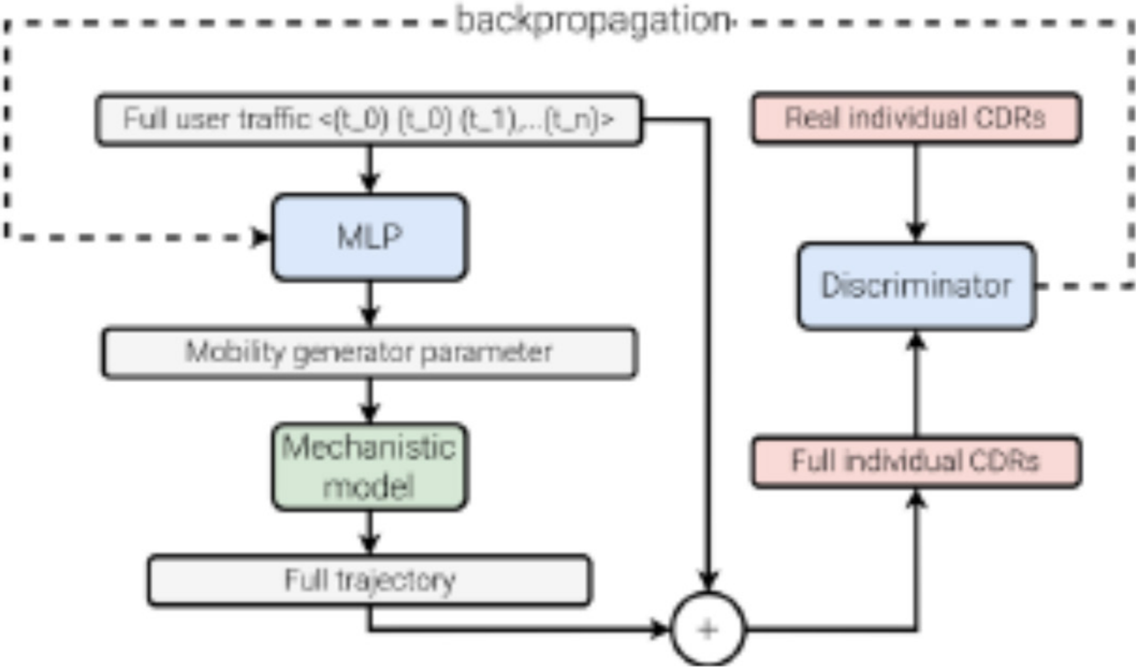


Figure 2: "Split-Joint" XDR generation model architecture

During two visit weeks of joint efforts towards this goal in CNR Pisa with Luca Pappalardo as a host, we ended up with a novel generative model with promising performance and evident beauty. As depicted in Figure 2, our modeling seamlessly combines both deep learning recent techniques highly performant in NLP applications (i.e., self-attention layers [R2]) and the literature legacy on human mobility laws and research ([R3], [R4]).

In particular, we focus on data traffic only (eXtended Data Records) and encode each user traffic as a sequence of the counts of her created data sessions per time slot of fixed length (e.g., 10 mins). Such segmentation allows the model to directly learn the circadian rhythm associated with human events generation, while the mapping of events to their exact timestamp can be done subsequently with a minimized error using an interpolation method, for instance. To cope with the high dimensionality of each user's sequence such modeling induces, we propose to leverage a self-attention layer instead of an LSTM to minimize the account of previous sequence elements corresponding to time slots further than an hour to the time slot of interest. Once the model is trained with only the traffic part of the input dataset, each of its produced users sequence is correlated to the process of generating a realistic trajectory in such a way that the output individual XDR (combined mobility and traffic) fits within the distribution of real users XDR sequences. The mobility trajectory is handled by a mechanistic model of the literature with a high fidelity of reproducing human laws in mobility and its flexibility to consider the mobility zone realistically. At last, the overall model training is done in an adversarial strategy.

In future steps, we plan to implement such modeling, starting with its traffic part. As the complete dataset will only be accessible in Chile for privacy compliance, we are organizing a mission there for this purpose. The mission goal is to train the model and validate the generation results in terms of comparisons of distributions with real-world ones and the general utility of the synthetic dataset in practical applications. This will require handling the raw dataset specificities, i.e., big size and attributes.

REFERENCES:

[R1] Anne Josiane Kouam, Aline Carneiro Viana, Alain Tchana. 2023. LSTM-based generation of cellular network traffic. IEEE WCNC 2023.

[R2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is all you need. 2017. Part of Advances in Neural Information Processing Systems 30 (NIPS 2017)

[R3] Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shouna Athavale, and Marta C. González. 2016. The TimeGeo modeling framework for urban mobility without travel surveys. PNAS. doi:10.1073/pnas.1524261113

[R4] Pappalardo, L., Simini, F. Data-driven generation of spatio-temporal routines in human mobility. Data Min Knowl Disc 32, 787–829 (2018). https://doi.org/10.1007/s10618-017-0548-4

Interpretable neural embedding on graphs

From a TNA experience by Simone Piaggesi

Learning latent low-dimensional vectors of network's nodes is the central aim of GRL [1], and nowadays node embeddings are crucial in order to solve machine learning tasks on graphs. Usually they are computed with self-supervised training, using edge reconstruction as a pretext task [2], with the result of encoding node proximities into distances of a metric space. According to this strategy, common to many algorithms such as DeepWalk and Node2Vec [2], inner products between node embedding pairs are proportional to the likelihood of observing the corresponding edge on the training graph.

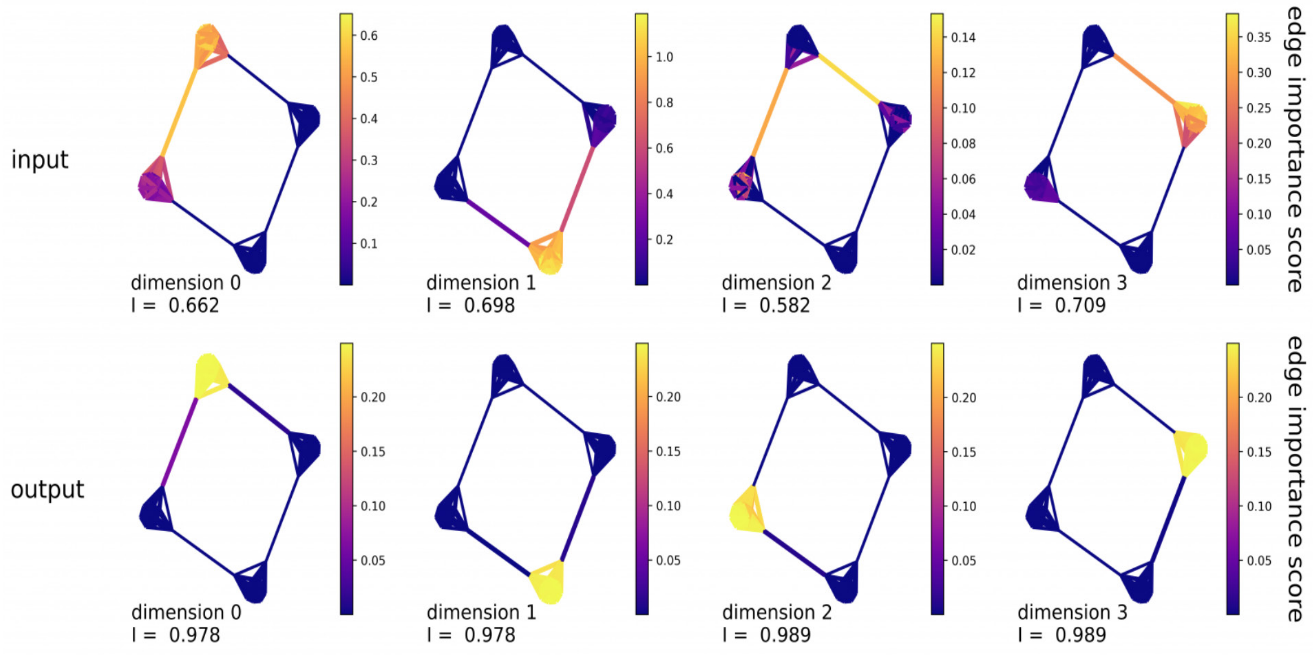


Figure 1: Edge importance scores computed with 4-dimensional embeddings

Node representations are understandable only in terms of their pairwise geometric relationships, and interpretations of individual embedding dimensions are typically hard to provide [3]. Here we define the interpretability of node embedding vectors assigning a human-understandable meaning to representation dimensions. In particular, we aim to associate different dimensions with different graph partitions [4]. In fact, densely connected subgraphs describe groups of nodes highly related to each other, like groups of similar words appear together when talking about a given topic [5].

We designed an algorithm that, given a set of node embeddings previously trained, returns a new set of post-processed node representations with a human-interpretable meaning. This is possible by means of autoencoder neural networks [6], which efficiently map input vectors into a new embedding space where we enforce interpretability constraints. In particular, we require orthogonality among clusters of edges reconstructed with different dimensions, in addition to calibrating new representations to ensure that all dimensions reconstruct a non-null subset of links.

In Figure 1 we see how the algorithm works on a toy graph with 4 densely connected communities arranged in a ring, i.e. a ring-of-cliques graph. We show, for any edge, variations of inner product scores (edge importance) occurring when removing the corresponding feature from 4-dimensional node vectors. As compared to input DeepWalk vectors, our technique leads to more interpretable edge reconstruction patterns on the output.

In fact, links from the same community receive high contributions by one unique dimension. Interpretability of dimensions is assessed by a score which quantifies how well the reconstructed edges match with a given partition.

In Figure 2 are shown the effects of our algorithm on the Cora citation network [7]. Nodes belonging to the same research topic are more likely to connect with each other, forming densely connected clusters. On the right (Top) we evaluate the accuracy [8] of interpretations, i.e. the level of alignment between representation dimensions and ground-truth topics. On the right (Bottom) we evaluate the fidelity [8] of interpretations, i.e. the link prediction agreement between the interpretable model and the input model. We observe, for our technique, increments in scores for interpretability (up to 80%) and link prediction (up to 10%), with respect to input DeepWalk vectors.

We have similar findings on multiple datasets, opening the path to new research investigations on the interpretability of unsupervised graph representation learning.

REFERENCES:

[1] Hamilton, William L. “Graph representation learning.” Synthesis Lectures on Artificial Intelligence and Machine Learning 14.3 (2020): 1-159.

[2] Hamilton, William L., Rex Ying, and Jure Leskovec. “Representation learning on graphs: Methods and applications.” arXiv preprint arXiv:1709.05584 (2017).

[3] Dalmia, Ayushi, and Manish Gupta. “Towards interpretation of node embeddings.” Companion Proceedings of the The Web Conference 2018. 2018.

[4] Girvan, Michelle, and Mark EJ Newman. “Community structure in social and biological networks.” Proceedings of the national academy of sciences 99.12 (2002): 7821-7826.

[5] Şenel, Lütfi Kerem, et al. “Semantic structure and interpretability of word embeddings.” IEEE/ACM Transactions on Audio, Speech, and Language Processing 26.10 (2018): 1769-1779.

[6] Baldi, Pierre. “Autoencoders, unsupervised learning, and deep architectures.” Proceedings of ICML workshop on unsupervised and transfer learning. JMLR Workshop and Conference Proceedings, 2012.

[7] Sen, Prithviraj, et al. “Collective classification in network data.” AI magazine 29.3 (2008): 93-93.

[8] Yuan, Hao, et al. “Explainability in graph neural networks: A taxonomic survey.” arXiv preprint arXiv:2012.15445 (2020).

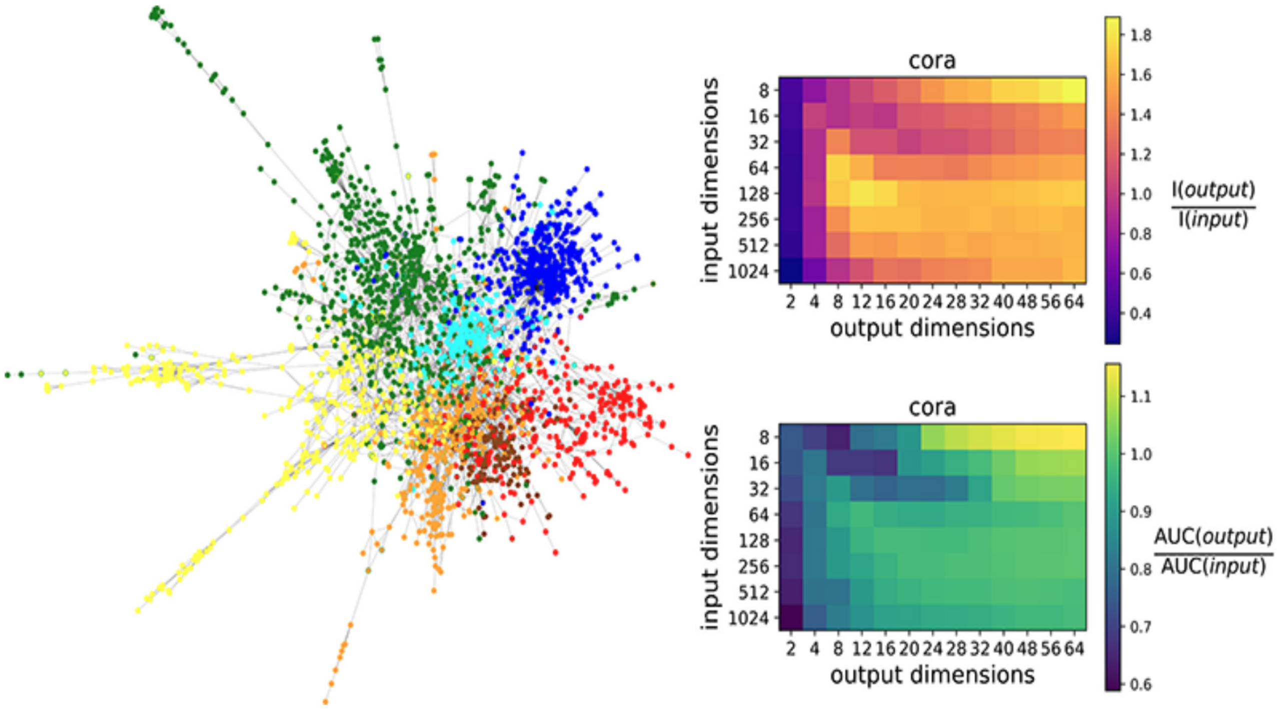


Figure 2: Accuracy and Fidelity gains between input DeepWalk and our output: (Left) visualization of Cora graph, image from https://arxiv.org/abs/1611.08402; (Top Right) ratios of average interpretability dimensions scores; (Bottom Right) ratios of average link prediction scores.

A TNA experience on Explainable AI

Dongwon Lee, Penn State University, USA

From June 1, 2022 to July 31, 2022, for two months, I had a privileged opportunity to visit the SoBigData++ team at Pisa, Italy, led by Prof. Fosca Gianotti at Scuola Normale Superiore (SNS) and Prof. Dino Pedreschi at University of Pisa. This was part of a program run by National Science Foundation (NSF), US that connects a PI in the US to a PI in the Europe with overlapping interests and partially supports the visit for research collaboration by paying airfares.

Once I got the award notice from NSF, I've searched the list of ERC projects that match my research interests and expertise, and found that the Explainable AI (XAI) project (https://xai-project.eu/) by Prof. Gianotti fit the bill exactly. For the last a few years, my research group has focused on the understanding, modeling, detection, and prevention of fake news in the US, and developed several cutting-edge detection algorithms and released benchmark datasets in the NSF-sponsored SysFake project (https://sites.google.com/site/pikesysfake/home). Despite accurate detection, however, our machine learning solutions have limited capabilities in explaining the verdict of fake news detection to other algorithms and human users. Therefore, the objectives of my visit to Prof. Gianotti's team were to learn the novel findings and methods from the XAI project, and seek for ways to apply them in my project.

During my 2-month-long visit, I made two research presentations: (1) Combating (Neural) False Information, AI & Society Summer School, University of Pisa, July 8, 2022, and (2) XAI for Non-Experts: Three Case Studies, Scuola Normale Superiore (SNS), June 8, 2022, and had also a chance to visit Dr. Mirco Nanni at ISTI - CNR in Pisa and discussed potential research collaboration ideas. Further, it turns out that the XAI project team was building a platform to collect and benchmark various XAI algorithms so that I donated the code of our XAI method (i.e., GRACE, KDD 2020: https://pike.psu.edu/publications/kdd20-grace.pdf). Finally, after numerous research meetings with Prof. Gianotti's team, we came up with two potential collaboration ideas: (1) Improving existing XAI methods using KG (knowledge graph) techniques that my group had some prior work, and (2) Applying XAI methods to the membership inference attack in security, which seems to be a novel usage of XAI. We hope to continue our research collaboration through Fall so that we can have some concrete outcomes.

In addition to research activities, I also had wonderful time visiting Rome, Florence, and Venice during weekends, and get to appreciate Italian culture and foods. Despite abnormally boiling weather in Italy this summer, it was truly wonderful experience. Below, the first photo shows me in front of the main office of SNS, Palazzo della Carovana that's built in 15th century, where my office was located. One could clearly see the famous Leaning Tower of Pisa through the office windows of Palazzo della Carovana. The second photo shows me and both Prof. Gianotti and Prof. Pedreschi, enjoying the sunset at a beach-front restaurant, after finishing the AI & Society summer school in July 2022.

I'd like to end this article by tremendously thanking my host, Prof. Gianotti, who has been very kind and generous, connecting me to the researchers and students in Pisa with related research interests, and helping me enjoy my stay in Italy. I feel quite indebted for the extra time and efforts that she spent in all administrative tasks and forms that she had to fill. I also acknowledge and express my gratitude to SoBigData++ program for supporting my visit.

Reducing exposure to harmful content in recommender systems

Corinna Coupette | Max Planck Institute for Informatics | from a TNA Experience at KTH Stockholm

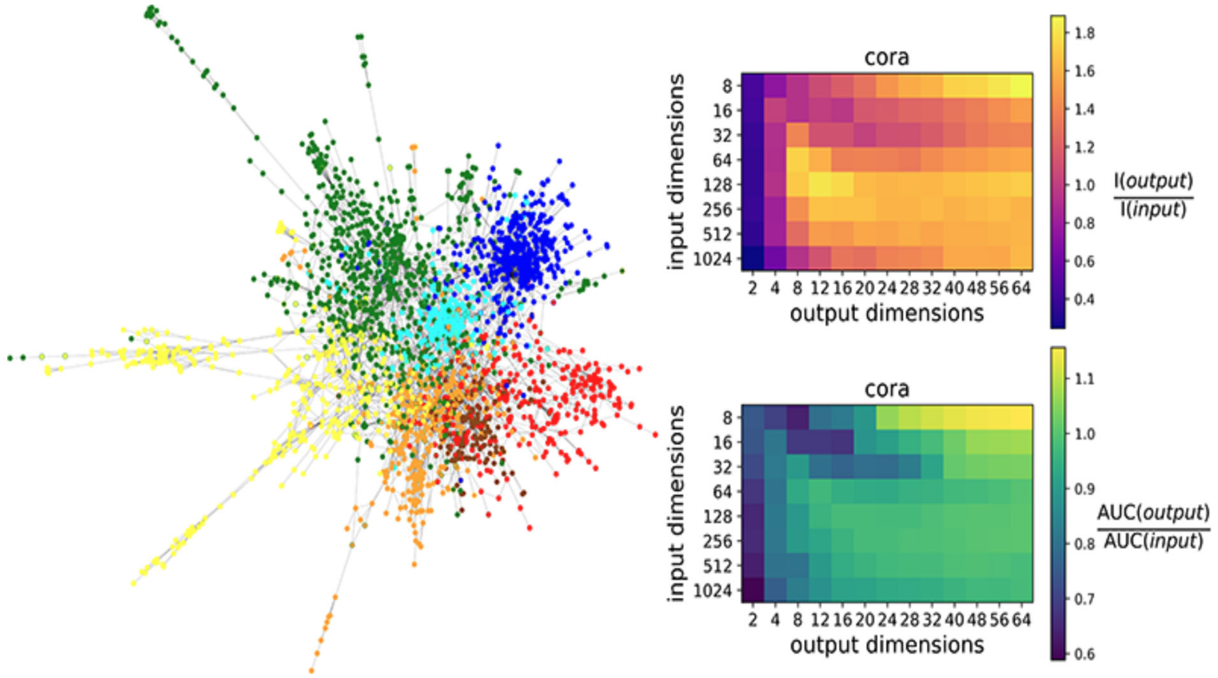


Figure 1: Modeling a digital media platform with some content labeled "harmful" as a directed graph with differently colored nodes. Dashed lines around a node indicate that its outgoing edges are not shown in the figure.

The recommender systems employed by digital media platforms typically showcase content that is similar to the content a user has already consumed. When that content is biased, this runs the risk of fuelling radicalization. One option to mitigate the radicalization risks posed by digital media platforms is to modify their recommendations. Motivated by empirical findings [3], recent work in the intersection of algorithm design and computational social science has made first strides toward formalizing this scenario as an optimization problem under budget constraints [1,2]. However, the resulting models, objective functions, and algorithms have a number of theoretical and practical drawbacks.

Our ongoing project, which was kick-started by the author's SoBigData++ TNA visit at KTH Stockholm, seeks to eliminate some of these limitations. It strives to provide a more realistic model and objective function for the recommendation rewiring problem, along with an efficient optimization algorithm to address the problem under the new model.

Thus far, we have developed a more natural formulation of the recommendation rewiring problem, capturing user behavior and anti-radicalization rewiring objectives more accurately than previous proposals. We have analyzed our model theoretically and proved that solving the resulting optimization problem is computationally hard in most settings (just like the less natural problem studied in [1]). We have started experiments with competing methods under our new model, and we are currently developing an efficient algorithm to optimize our new objective function.

Like inspirational prior work [1], we model a digital media platform as a directed graph in which nodes represent content items, directed edges represent recommendations, and each content item is assigned a label indicating whether it is considered harmful (Figure 1). We also focus on edge rewiring as our main graph operation to improve our objective function, which intuitively corresponds to replacing a recommended item by another one (e.g., when watching YouTube video A, the user now gets recommended video C, rather than video B). However, we model users more realistically than prior work, and our optimization objective assesses the reachability of harmful content much more holistically.

While the formalizations of the rewiring problem studied in past work were relatively easy to analyze but unnatural, more natural formalizations are harder to analyze, and consequently, harder to tackle with algorithms that also provide guarantees.

As the potential of recommendation rewiring to reduce harmful content exposure on digital media platforms heavily depends on the chosen formalization, more work developing analytical techniques and approximation algorithms for such formalizations is desirable. In our own work, completing the project started during the author's SoBigData++ TNA, we hope to provide a theoretically sound and scalable approximation algorithm for the most natural formalization of the rewiring problem studied to date. Beyond this immediate next step, we see two promising directions for future work. First, on the theoretical side, we currently lack a comprehensive overview of hardness results for the rewiring problem and its variants, which would be very useful for algorithm design. Second, on the applied side, a framework supporting the development of theoretically sound approximation algorithms for variants of the rewiring problem would be highly valuable in practice.

REFERENCES

[1] Fabbri, Francesco, Yanhao Wang, Francesco Bonchi, Carlos Castillo, and Michael Mathioudakis. "Rewiring What-to-Watch-Next Recommendations to Reduce Radicalization Pathways." In Proceedings of the ACM Web Conference 2022, pp. 2719-2728. 2022.

[2] Haddadan, Shahrzad, Cristina Menghini, Matteo Riondato, and Eli Upfal. "Repbublik: Reducing polarized bubble radius with link insertions." In Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp. 139-147. 2021.

[3] Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. "Auditing radicalization pathways on YouTube." In Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 131-141. 2020.

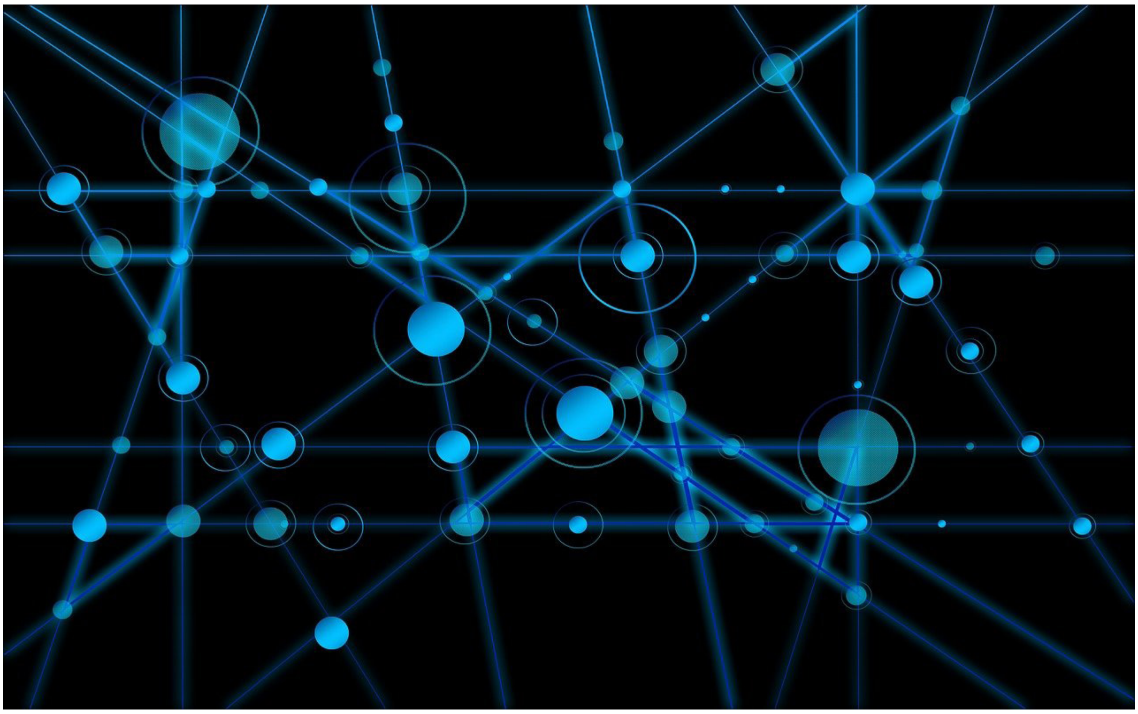


Image by Gerd Altmann from Pixabay

Algorithms for characterizing the spreading of misinformation on social networks through the lens of temporal networks

Ilie Sarpe | University of Padova, Italy | TNA expericence at KTH Royal Institute of Technology in Stockholm.

Social networks enable us to stay in contact with people all over the world, and to read news related to events that occur over the globe in almost real-time. While this has many positive impacts, such as allowing the report of adverse events. Unfortunately, such speed of diffusion of information over social networks has also important negative implications. In many cases malicious users may try to take advantage of the structure of social networks to spread misinformation. Such news can be really convincing that even user without bad intents may end-up sharing and spreading misinformation. To contrast such aspect, it is of crucial importance to understand how misinformation spreads over social networks, and in particular identifying all those nodes that systematically spread misinformation over social networks.

A NEW CHALLENGE

In the current big data era, content over social networks is produced at an unprecedent rate. While companies that rule such complex systems put significant effort in contrasting the spreading of misinformation, many users get exposed to false or misleading information. This has a practical impact in the current society, since misinformation can be used to bias or polarize opinions on many controversial themes such COVID-19 or political elections, that have practical impacts on everyday lives. Many approaches have been proposed to contrast the spread of misinformation, such as for example text-based approaches that aim at flagging misinformation based on text processing. While this and other approaches can

Image by Gerd Altmann from Pixabay

filter some misinformative content, some text cannot be filtered since it appears fact-checked, e.g., contents may intentionally distort reality even quoting scientific references. Therefore, novel techniques that consider the underlying patterns of the spreading process carried by the users, to complement the existing approaches are needed.

SOCIAL NETWORKS AS TEMPORAL NETWORKS AND PATTERNS

Social networks can be modeled as graphs or networks and many algorithmic approaches can be used on such model to infer desired properties of the network. Unfortunately, usual (static) networks do not account for the timing of occurrence of the events in a network. The timings of occurrence of event in fact are related to how information spreads over a network, for example if three users are connected by a path (i.e., there is a flow of information spreading from the first to the last user through the middle user) but the timing of the first edge on the path is greater than the timing on the second edge on the path then information cannot flow on such path. Accounting for timing of events can provide us a new lens for identifying important properties of the social networks modeled through temporal networks. Temporal patterns are defined as frequent structures that occur repeatedly in a short amount of time over a temporal network, these patterns capture both the way users interact over the network and the dynamics through which such interactions are performed. By analyzing different patterns, we are able to understand if a user behavior is frequent or infrequent and how it compares to the different patterns tested. Therefore, through such analysis we are able to correlate users with important functions over the social network.

TEMPORAL PATTERNS AND ALGORITHMS TO TACKLE MISINFORMATION

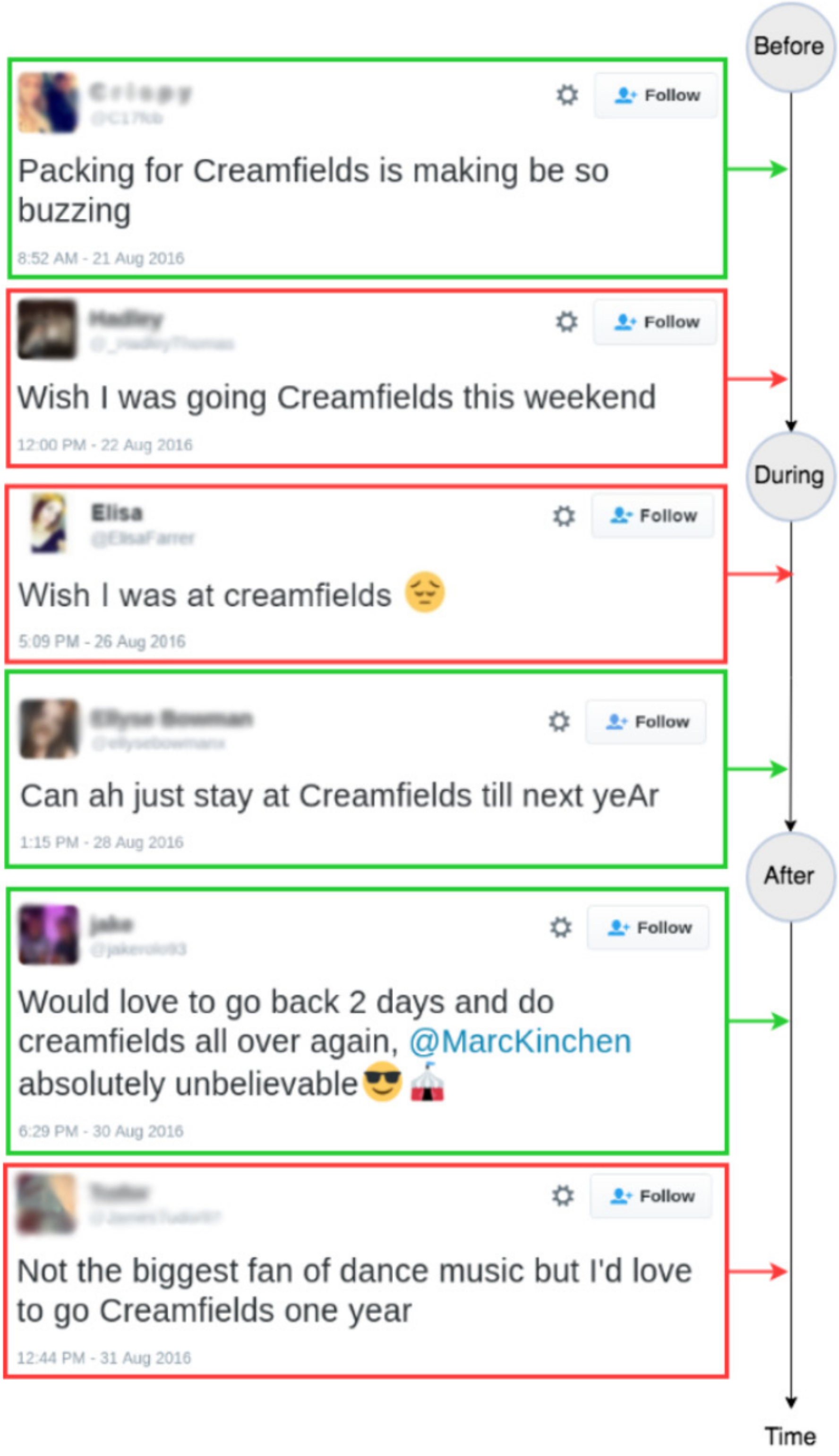
In my experience at KTH, working with Professor Aristides Gionis, I had the possibility to model the problem of detecting how misinformation spreads across social networks with temporal networks and temporal patters. Since a temporal pattern can encode several ways of spreading misinformation over a social network, we worked on the problem of identifying those users that most contribute to realizing such patterns. This is a novel problem, never tackled in literature on temporal networks and it has several challenges, including scaling the computation on billion edges networks, and providing rigorous theoretical guarantees on the quality of the solution computed. We were able to address this problem and to provide new algorithmics tool to identify nodes that contribute to a high percentage of actions that lead to the spreading of misinformation across social networks modeled as temporal networks. We provide rigorous algorithms to address this problem and we believe that such algorithms will have a practical impact in many scenarios, since we also addressed the problems of scalability and efficiency mentioned previously.

SUMMARIZING

Thanks to the SoBigData++ TNA fellowship I had the possibility to visit the research group of Professor Aristides Gionis and carry on a research project in a very stimulating environment. At KTH I met many researchers working on topics related to my research interests and had the possibility to share and discuss many ideas. Working with Prof. Gionis we developed several algorithms to detect the spread of misinformation by leveraging on temporal patterns, we also think that such algorithms will be of practical impact in many scenarios given that such problem was never addressed given its complexity.

Event attendance prediction using social media

Cristina Muntean, ISTI-CNR



Packing for Creamfields is making be so buzzing
8:52 AM - 21 Aug 2016

Wish I was going Creamfields this weekend
12:00 PM - 22 Aug 2016

Wish I was at creamfields 😔
5:09 PM - 26 Aug 2016

Can ah just stay at Creamfields till next yeAr
1:15 PM - 28 Aug 2016

Would love to go back 2 days and do creamfields all over again, @MarcKinchen absolutely unbelievable 😎 🎪
6:29 PM - 30 Aug 2016

Not the biggest fan of dance music but I'd love to go Creamfields one year
12:44 PM - 31 Aug 2016

Fig. 1. Examples of tweets posted before, during and after the event

Popular social media applications on smartphones (e.g. Facebook, Instagram, Twitter) enabled the creation of an unprecedented amount of user-generated content. Social media can be useful to extract valuable information concerning human dynamics and behaviors, such as mobility. Popular events such as music festivals attract thousands of participants. Usually, the presence is well reflected in social media networks, allowing people to connect with "the event", expressing through posts their feelings, experiences or opinions well in advance of its planned date.

Given the attention to popular events reflected in social media, we tackle a novel, interesting problem: Is it possible to infer from Twitter posts the actual attendance of the user to the cited event? To answer this question, we conducted experiments on data from two large music festivals in the UK, namely the VFestival and Creamfields events. The research has been published in two important venues, a preliminary study at the ASONAM conference (ref.1) and a more detailed study in the IPM journal (ref.2).

The simplest way of inferring the users' presence at events is to consider the geotag associated with their posts: the "check-in" or the user location in the event place at the time of the event can trivially be associated with attendance. There are nevertheless two drawbacks to this approach. The first drawback is that few social media users enable the geotagging of their posts (on Twitter, the percentage of geotagged posts is about 2%). Using this data to learn attendance prediction classifiers would be difficult and may lead to ineffective predictive models due to its sparsity. The second drawback of only using geolocated data is that they do not represent the intention of the user to participate in the event. To avoid these two aforementioned issues, we wish to infer the actual attendance of users to an event by only relying on the content of non-geotagged posts, without considering any spatial features.

For the event attendance classification task, we devise three distinct temporal intervals identifying when the posts have been shared on social media: before, during or after the event. For each of these three, we propose distinct classification tasks. The analysis of posts shared before the event serve as a predictor of the users' actual attendance, the analysis of posts shared during the event reflects the actual participation of users at the event, while the analysis of posts shared after the event offers an overview of past attendance. We come up with four different categories of features. Each category reflects a different facet of social media, namely the: textual, temporal, social, and multimedia dimensions.

Particularly interesting is the "before" case, since an early knowledge of the possible user attendance can be useful for proposing innovative services and applications. For example, event organizers or third-party companies could precisely target their advertisement campaigns by offering personalized services to the users most probable to participate in the event. Another relevant example is transportation planning, where attendance prediction could allow the organizers or the local authorities to urge potential attendees to use public transportation or can help bus and shuttle companies to plan and advertise collective transport services to the event.

The created models achieve a very high accuracy with the highest result observed for the Creamfields festival, exhibiting ~91% accuracy at classifying users that have expressed their intention to attend the event. Some of the most prominent features are the word embedding features that contribute to achieving high performance. The analysis of visual content is a growing trend in social media and could be successfully explored in the classification process through the use of deep learning techniques.

REFERENCES:

Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, Valéria Cesário Times: Exploring Social Media for Event Attendance. ASONAM 2017: 447-450

Vinicius Monteiro de Lira, Craig Macdonald, Iadh Ounis, Raffaele Perego, Chiara Renso, Valéria Cesário Times: Event attendance classification in social media. Inf. Process. Manage. 56(3): 687-703 (2019)

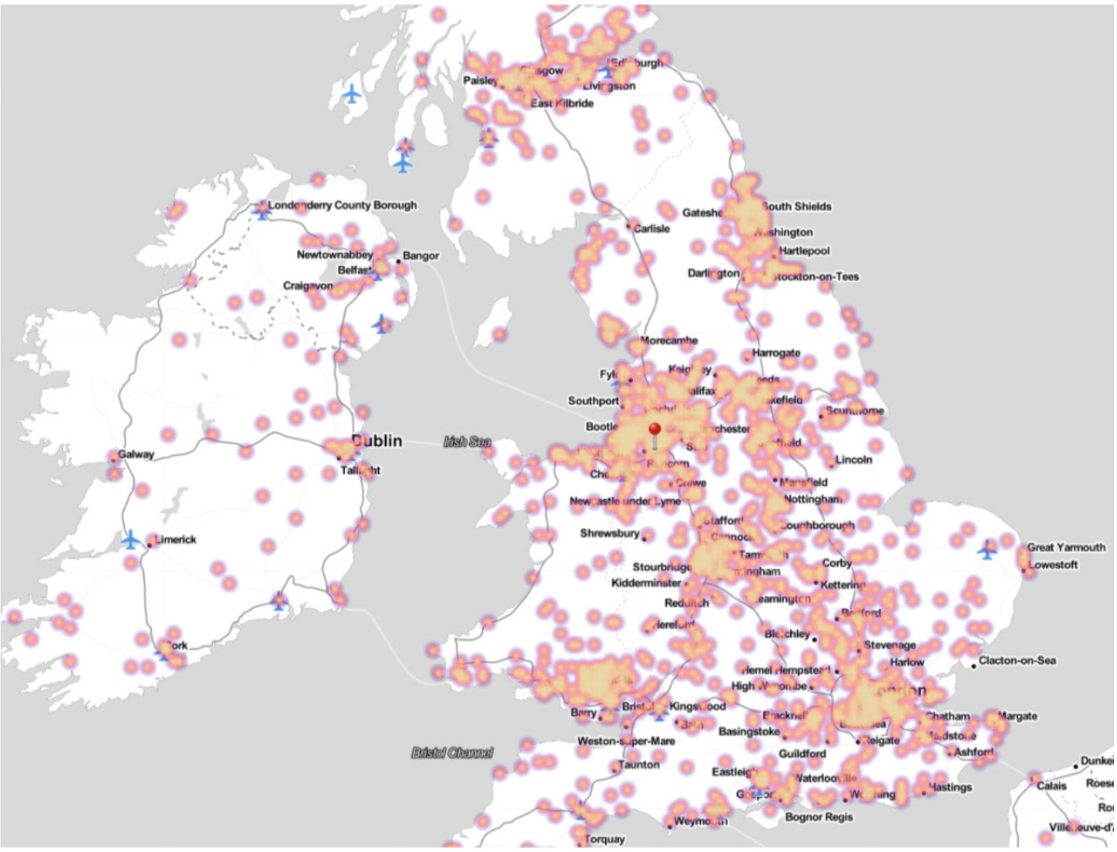


Fig. 2. Heatmap with distribution by hometown of the inferred attendees at the Creamfields festival (red point).

Load ready-to-use mobility datasets with one line of code!

Giuliano Cornacchia, Ph.D. Student in Computer Science, University of Pisa & ISTI-CNR

The new version of scikit-mobility includes a "data module" to access and download ready-to-use mobility datasets and upload new ones to the collection!

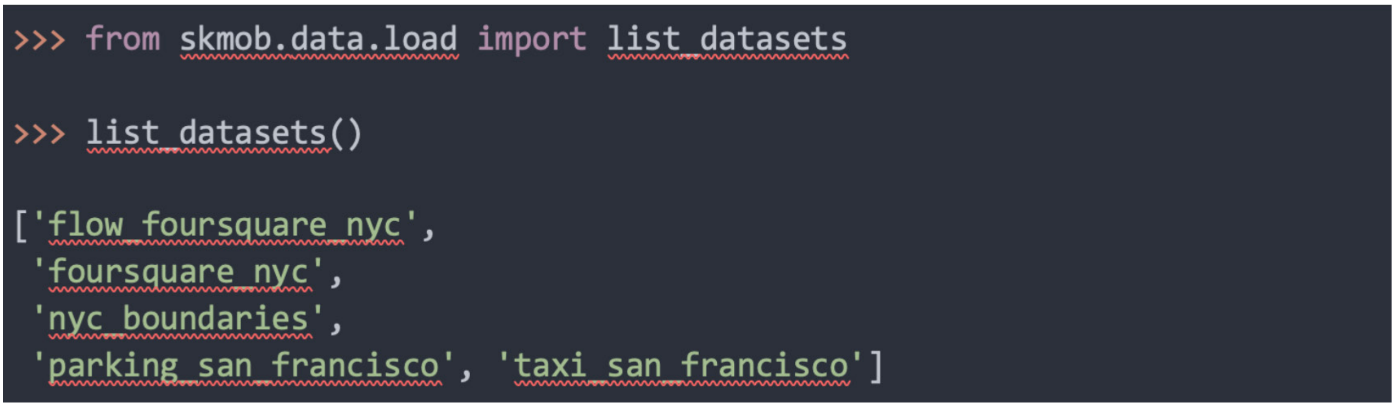
As data scientists, we all know the convenience of having ready-to-use datasets at your fingertips. When performing exploratory data analysis (EDA), testing hypotheses, and prototyping models, nothing can beat having preprocessed and reliable datasets for experimentation.

Unfortunately, while there are widely known datasets for tasks such as digit recognition (MNIST), classification (IRIS), and sentiment analysis (Tweets), this is not the case in the human mobility domain.

That's why we, scikit-mobility's developers, decided to fill this gap by providing a module to access and download custom-curated mobility datasets, spanning from GPS traces to Origin-Destination matrices and everything in between.

Scikit-mobility (skmob) is among the most used python libraries for human mobility analysis, comprising modules for preprocessing, synthetic traces generation, trajectory data mining, and flow estimation.

Within the new skmob's release, you can find the data module, which allows you to retrieve standard benchmarking datasets via an easy-to-use interface. The data module API consists of two main functions: list_datasets and load_dataset. The former shows the datasets already available

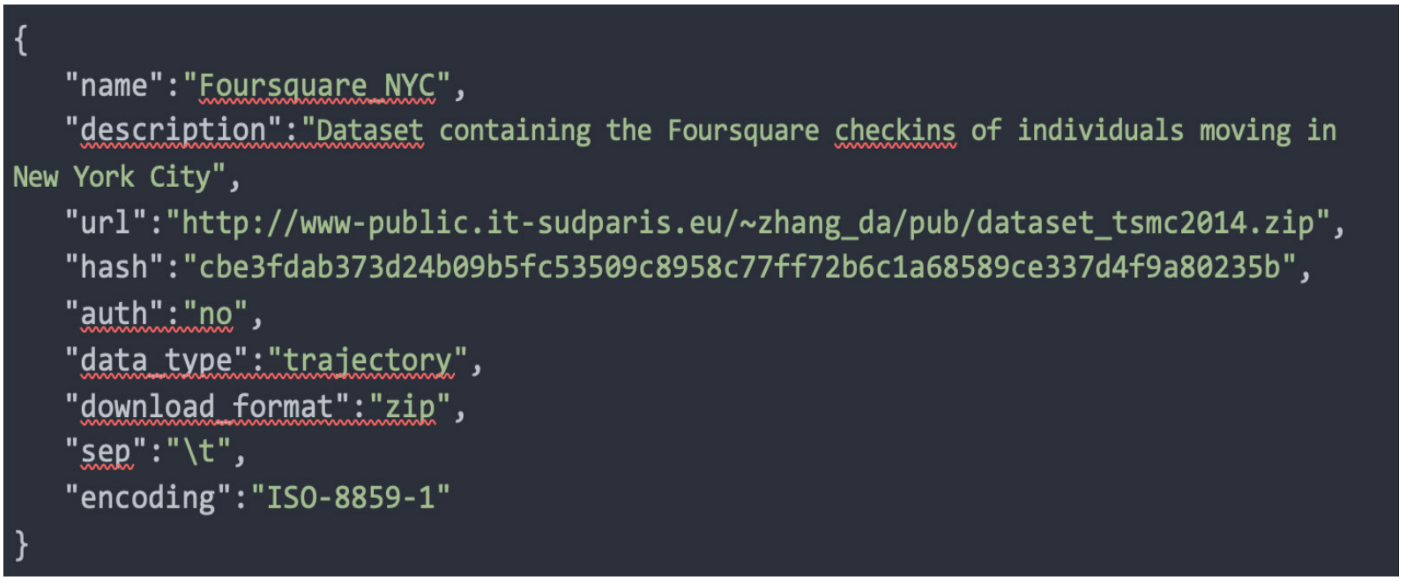


```
>>> from skmob.data.load import list_datasets

>>> list_datasets()

['flow_foursquare_nyc',
 'foursquare_nyc',
 'nyc_boundaries',
 'parking_san_francisco', 'taxi_san_francisco']
```

The python code required to list all the available dataset



```
{
   "name":"Foursquare_NYC",
   "description":"Dataset containing the Foursquare checkins of individuals moving in
New York City",
   "url":"http://www-public.it-sudparis.eu/~zhang_da/pub/dataset_tsmc2014.zip",
   "hash":"cbe3fdab373d24b09b5fc53509c8958c77ff72b6c1a68589ce337d4f9a80235b",
   "auth":"no",
   "data_type":"trajectory",
   "download_format":"zip",
   "sep":"\t",
   "encoding":"ISO-8859-1"
}
```

The JSON manifest for the Foursquare NYC dataset in the repository, the latter retrieves the requested dataset and directly outputs it into a skmob-friendly data structure.

Furthermore, the data module lets you contribute to the human mobility community by uploading your own datasets and making them available for researchers, industry practitioners, and academia at large. You only need to upload it somewhere publicly accessible on the internet (e.g., through an URL).

The way to expose your uploaded dataset is straightforward, and it is done via a JSON manifest file and a pre-processing python function:

- The JSON manifest describes all the relevant metadata necessary to retrieve your dataset, e.g., the URL at which the dataset is available, name, description, license, maintainers, and citation (if required).

- The pre-processing function contains all the pre-processing steps and data transformations necessary to make your dataset compliant with skmob data structures.

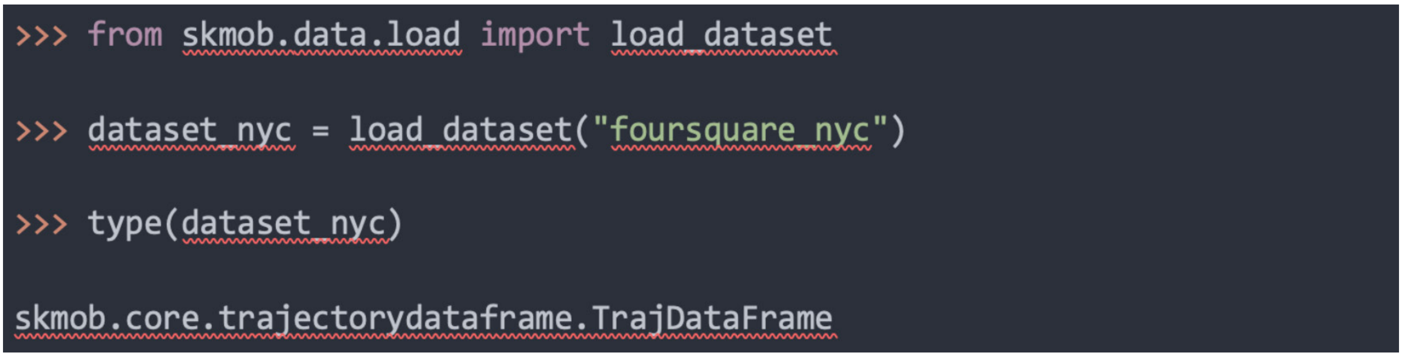
To consume an uploaded dataset, a user simply has to write only ONE line of code calling the load_dataset function with the name of the dataset the user wants to download. This function downloads, pre-processes and transforms your data to ensure scikit-mobility data standards, and make it available for analysis!

Are you interested in contributing?

You may contribute and maybe upload the "IRIS" dataset for mobility! Reach as out @scikitmobility

Data module video tutorial: https://youtu.be/FjJZsaHHuvw

Example notebook: https://jovian.ai/giuliano-cornacchia/the-data-module



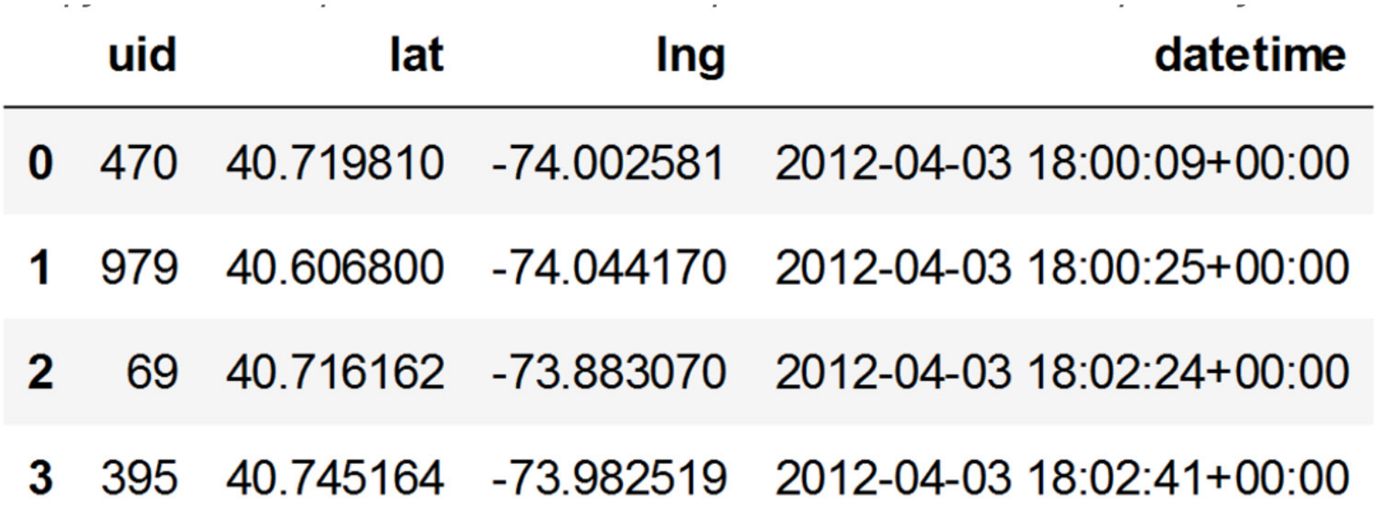
```
>>> from skmob.data.load import load_dataset

>>> dataset_nyc = load_dataset("foursquare_nyc")

>>> type(dataset_nyc)

skmob.core.trajectorydataframe.TrajDataFrame
```

The python code required to load the Foursquare NYC dataset (foursquare_nyc)



	uid	lat	lng	datetime
0	470	40.719810	-74.002581	2012-04-03 18:00:09+00:00
1	979	40.606800	-74.044170	2012-04-03 18:00:25+00:00
2	69	40.716162	-73.883070	2012-04-03 18:02:24+00:00
3	395	40.745164	-73.982519	2012-04-03 18:02:41+00:00

A sample of the Foursquare NYC dataset (foursquare_nyc)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

SoBigData Magazine is published under the

project N° 871042 | Programme: H2020 - INFRAIA
Duration: 01/01/2020 - 31/12/2024

Editorial Secretariat

info@sobigdata.eu

Layout and Design

Beatrice Rapisarda

Privacy statement

The personal data (names, email addresses...) and the other information entered in SoBigData Magazine will be treated according with the provision set out in Legislative Degree 196/2003 (known as Privacy Code) and subsequently integration and amendement.

Coordinator and Legal representative of the project: Roberto Trasarti | roberto.trasarti@isti.cnr.it

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

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

SoBigData

SoBigData

www.sobigdata.eu