Gabriele Tolomei, Ph.D.

email: tolomei@di.uniroma1.it ⋄ **home:** http://www.di.uniroma1.it/~tolomei/
ORCID ID: orcid.org/0000-0001-7471-6659

EDUCATION

National Scientific Qualification (ASN)
Role: Professore di Seconda Fascia (Associate Professor)
Settore Concorsuale (Academic Field): 01/B1 - Informatica (Informatics)
From - To: 7 August 2018 - 7 August 2024
Role: Professore di Seconda Fascia (Associate Professor)
Settore Concorsuale (Academic Field): 09/H1 - Sistemi di Elaborazione delle Informazioni (Information Processing Systems)
From - To: 26 July 2018 - 26 July 2024

Ph.D. in Computer Science *01/2008 - 11/2011*
Università Ca' Foscari Venezia, Italy
Date: 17 November 2011
Thesis Title: Enhancing web search user experience: from document retrieval to task recommendation
Supervisors: Salvatore Orlando and Fabrizio Silvestri
Main Results: Developed an algorithm to discover the set of *user tasks* (i.e., group of search queries having the same latent need) from historical data stored in search engine logs. This solution performed 16% better than traditional techniques in terms of F1 score, and about 5% better than the very best state-of-the-art method known at that time.
Most valuable results published in ACM WSDM 2011 (best paper runner up) and ACM TOIS.

M.Sc. in Computer Science (*summa cum laude*) *10/2002 - 04/2005*
Università di Pisa, Italy
Date: 21 April 2005

B.Sc. in Computer Science *10/1999 - 10/2002*
Università di Pisa, Italy
Date: 18 October 2002

RESEARCH EXPERIENCE

Associate Professor *09/2019 -*
Sapienza University of Rome, Italy

Assistant Professor *07/2017 - 08/2019*
Università degli Studi di Padova, Italy
From - To: 18 July 2017 -
Goals: Research activities on topics at the intersection of machine learning and computer security. Establishment of a multidisciplinary team focused on *adversarial machine learning* in collaboration with Università Ca' Foscari di Venezia, Italy.

Projects:

- *Interpretability of machine learning models*: Formulated the problem of finding the "best" (i.e., less costly) perturbations of input features so as to switch the predictions output by an existing tree-based ensemble classifier. An algorithm to solve the problem has been proposed and its validity has been assessed on a real-world use case (i.e., online advertising). Results have been published in ACM KDD 2017 conference; an extended manuscript has been submitted to IEEE TKDE journal, and it is about to be submitted for second-round review.

- *Robustness of machine learning models*: Definition of the problem of training machine learning models that are insensitive to (i.e., robust against) input perturbations crafted by a malicious attacker, inspired by the notion of *non-interference* that is typical of the computer security domain. Proposal of a solution which is validated on public datasets. Results have been and will be submitted for review to the ACM CIKM 2019 and IEEE ICDE 2020 conferences, respectively.
- *CSRF attacks detection using machine learning*: CSRF attacks are one of the main web security threats. Supervised learning techniques have been used to train a prediction model (i.e., a binary classifier) on a dataset of labeled HTTP requests, collected with a browser extension developed *ad hoc*. The classifier outperforms any (heuristic-based) baselines, scoring $F_1 = 0.72$. Results and dataset have been published to the IEEE EuroS&P 2019 conference.
- *IoT advertising*: Online advertising is possibly the most profitable Internet-based business model yet it is still "limited" to traditional devices (i.e., PCs and smartphones). A new idea of advertising has been sketched so as to extend Internet advertising business to emerging pervasive and ubiquitous interconnected *smart devices*, which are collectively known as the *Internet of Things* (IoT). Such a novel vision – along with the challenges to be addressed – are described in a manuscript which appears in the IEEE Communications Magazine.
- *Fraud-free, verifiable advertising costs*: Ongoing collaboration with the Bosch Research and Technology Center of Pittsburgh, PA, USA. This project aims to introduce a new model of online advertising, which allows advertisers – who are often victims of frauds (e.g., ad click inflation) – to verify the amount of money they spend on their campaigns charged by ad networks and publishers.

Research Scientist *06/2014 - 07/2017*
Yahoo Labs, London, UK
From - To: 2 June 2014 - 14 July 2017
Goals: Improve the engagement of users with *Gemini*, the integrated Yahoo online advertising platform. Promote "high quality" advertisements using measures of post-click satisfaction, which go beyond traditional Click-Through Rate (CTR). Analyse large-scale datasets from distributed computing environments and mine interesting patterns via statistical/machine learning solutions. Design, implement, and test innovative prototypes into production buckets as response to internal challenges, in collaboration with Product and Engineering teams. Delivery results both internally and externally (e.g., research paper submissions to top conferences like SIGIR, KDD, CIKM, RecSys, WSDM, WWW, etc.).

Projects:

- *Accidental ad click discovery and discounting*: Design and implementation of a data-driven methodology to detect "accidental" clicks on CPC advertisements shown on Yahoo's properties, currently used in production. Proposal of a technique for discounting those clicks so to balance between inevitable drop in revenue and long-term satisfaction of advertisers. Proposed solution has been published to the International Journal of Data Science and Analytics (JDSA) and patented with the US Patent and Trademark Office.
- *Ad quality score*: Design and implementation of a mechanism to monitor and report to the advertisers the performance of their ad campaigns running on the Yahoo Gemini platform. Successfully tested on a pool of selected advertisers and patented with the US Patent and Trademark Office.
- *Ad feature recommendations*: Design and implementation of a system which is able to suggest actionable changes to ad landing pages so as to improve their *quality* perceived by users. The approach has been published to the ACM KDD 2017 conference and patented with the US Patent and Trademark Office.

Postdoctoral Research Fellow *01/2012 - 06/2014*
Università Ca' Foscari Venezia, Italy
From - To: 13 January 2012 - 1 June 2014

Goals: Novel application of machine learning and data mining techniques to large scale, heterogeneous data sources with the aim of improving the effectiveness of web search engines.

Projects:

- *Classification of web authentication cookies*: Automatic discovery of *authentication cookies* from those stored in web browsers using a supervised learning technique. Design of a (semi-) automatic method to build a ground truth of authentication cookies. Evaluation of four state-of-the-art solutions proposed to detect authentication cookies. Development of a binary classifier, which outperforms existing solutions by increasing the overall F_1 score from 14% up to 23%.
- *Trending topics vs. web search*: Analysed the impact of Twitter *trending entities* on user search behaviour. Time-series regression revealed that signals from Twitter are useful to predict Google Hot Trends and Wikipedia page requests or edits about 60% of times. Results published in CIKM 2013 workshop, ASE/IEEE SocialCom 2013.
- *Task-oriented web search and recommendation*: Developed a graph-based model of *task-based* user search behaviour. Implemented the prototype of a *task recommender system*, which suggested tasks to web users instead of "traditional" queries, with about 50% precision. Results published in OAIR 2013 conference and the ACM TOIS journal (**ACM 2013 Computing Reviews Notable Article**).

Research Assistant *01/2008 - 01/2012*
ISTI-CNR, Pisa, Italy
From - To: 16 January 2008 - 8 January 2012
Goals: Research on high performance computing with application to web search and mining.

Projects:

- *Task-oriented web search*: Developed an algorithm to discover *user tasks* (i.e., group of queries having the same latent need) from search engine logs. Evaluated F_1 score 16% better than traditional techniques, and about 5% better than the very best method known at that time.
  Results published in ACM WSDM 2011 conference (**best paper runner up**).

SELECTED PUBLICATIONS

Journals[1]

- Calzavara, S., Lucchese, C., Tolomei, G.. Abebe, S., and Orlando, S. *Treant: Training Evasion-Aware Decision Trees*. In Data Mining and Knowledge Discovery, 34(5): 1390-1420 (2020) [**impact factor = 2.629**].
- Calzavara, S., Conti, M., Focardi, R., Rabitti, A. and Tolomei, G. *Machine Learning for Web Vulnerability Detection: The Case of Cross-Site Request Forgery*. In IEEE Security & Privacy, 18(3): 8–16 (2020) [**impact factor = 1.596**].
- Tolomei, G. and Silvestri, F. *Generating Actionable Interpretations from Ensembles of Decision Trees*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), *in press* [**impact factor = 3.857**].
- Tolomei, G., Lalmas, M., Farahat, A., and Haines A. *You Must Have Clicked on this Ad by Mistake! Data-Driven Identification of Accidental Clicks on Mobile Ads with Applications to Advertiser Cost Discounting and Click-Through Rate Prediction*. In International Journal of Data Science and Analytics, Vol. 7, Issue 1, pp. 53–66.

[1] impact factor JCR 2017 (when available).

- Aksu, H., Babun, L., Conti, M., Tolomei, G., and Uluagac, A. S. *Advertising in the IoT Era: Vision and Challenges.* In IEEE Communications Magazine, Vol. 56, Issue 11, pp. 138–144 [**impact factor = 10.435**].

- Calzavara, S., Tolomei, G., Bugliesi, M., and Orlando, S. *A Supervised Learning Approach to Protect Client Authentication on the Web.* In ACM Transactions on the Web (TWEB), Vol. 9, Issue 3 - June 2015, Article No. 15, pp. 1–30 [**impact factor = 1.526**].

- Giummolè, F., Orlando, S., and Tolomei, G. *A Study on Microblog and Search Engine User Behaviors: How Twitter Trending Topics Help Predict Google Hot Queries.* In ASE Human Journal, Vol. 2, Issue 3 - September 2013, pp. 195–209.

- Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G. *Discovering Tasks from Search Engine Query Logs.* In ACM Transactions on Information Systems (TOIS), Vol. 31, Issue 3 - July 2013, pp. 1–43 [**impact factor = 2.312; ACM 2013 Computing Reviews Notable Article**[2]]

- Miori, V., Tarrini, L., Manca, M., and Tolomei, G. *An Open Standard Solution for Domotic Interoperability.* In IEEE Transactions on Consumer Electronics, Vol. 52, Issue 1 - February 2006, pp. 97–103 [**impact factor = 1.694**].

Conferences and Workshops[3]

- Calzavara, S., Lucchese, C., and Tolomei, G. *Adversarial Training of Gradient-Boosted Decision Trees.* In Proc. of ACM CIKM 2019, pp. 2429–2432 [**rank = A**].

- Calzavara, S., Conti, M., Focardi, R., Rabitti, A., and Tolomei, G. *Mitch: A Machine Learning Approach to the Black-Box Detection of CSRF Vulnerabilities.* In Proc. of IEEE Euro S&P 2019, pp. 528–543.

- Conti, M., Gangwal, A., Gochhayat, S. P., and Tolomei, G. *Spot the Difference: Your Bucket is Leaking : A Novel Methodology to Expose A/B Testing Effortlessly.* In Proc. of IEEE CNS 2018, pp. 1–7.

- Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. *Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking.* In Proc. of ACM KDD 2017, pp. 465–474 [**rank = A***].

- Lucchese, C., Nardini, F. M., Orlando, S., and Tolomei, G. *Learning to Rank User Queries to Detect Search Tasks.* In Proc. of ACM ICTIR 2016, pp. 157–166.

- Lalmas, M., Lehmann, J., Shaked, G., Silvestri, F., and Tolomei, G. *Promoting Positive Post-Click Experience for In-Stream Yahoo Gemini Users.* In Proc. of ACM KDD 2015, pp. 1929–1938 [**rank = A***].

- Calzavara, S., Tolomei, G., Bugliesi, M., and Orlando, S. *Quite a Mess in My Cookie Jar! Leveraging Machine Learning to Protect Web Authentication.* In Proc. of WWW 2014, pp. 189–200 [**rank = A***].

- Giummolè, F., Orlando, S., and Tolomei, G. *Trending Topics on Twitter Improve the Prediction of Google Hot Queries.* In Proc. of ASE/IEEE SocialCom 2013, pp. 39–44 [**rank = B; among the top-5% best papers**].

- Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G. *Modeling and Predicting the Task-by-Task Behavior of Search Engine Users.* In Proc. of OAIR 2013, pp. 77–84.

[2] http://www.computingreviews.com/recommend/bestof/notableitems_2013.cfm

[3] ranking CORE 2018 [http://portal.core.edu.au/conf-ranks/] (when available).

- Orlando, S., Pizzolon, F., and Tolomei, G. *SEED: A Framework for Extracting Social Events from Press News*. In Proc. of WWW-WoLE 2013, pp. 1285–1294 [**rank = A***].
- Ferrari, A., Gnesi, S., and Tolomei, G. *Using Clustering to Improve the Structure of Natural Language Requirements Documents*. In Proc. of REFSQ 2013, pp. 34–49 [**rank = B; best paper runner-up**].
- Bruni, E., Ferrari, A., Seyff, N., and Tolomei, G. *Automatic Analysis of Multimodal Requirements: A Research Preview*. In Proc. of REFSQ 2012, pp. 218–224 [**rank = B**].
- Ferrari, A., Gnesi, S., and Tolomei, G. *A clustering-based approach for discovering flaws in requirements specifications*. In Proc. of ACM SAC 2012, pp. 1043–1050 [**rank = B**].
- Ceccarelli, D., Gordea, S., Lucchese, C., Nardini, F.M., and Tolomei, G. *Improving Europeana Search Experience Using Query Logs*. In Proc. of TPDL 2011, pp. 384–395 [**rank = B**].
- Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G. *Identifying Task-based Sessions in Search Engine Query Logs*. In Proc. of ACM WSDM 2011, pp. 277–286 [**rank = A*; best paper runner-up**].
- Lucchese, C., Orlando, S., Perego, R., Silvestri, F., and Tolomei, G. *Detecting Task-based Query Sessions using Collaborative Knowledge*. In Proc. of WI-IAT 2010, pp. 128–131 [**rank = B**].
- Tolomei, G., Orlando, S., and Silvestri, F. *Towards a Task-based Search and Recommender Systems*. In Proc. of IEEE ICDE 2010, pp. 333–336 [**rank = A***].
- Mordacchini, M., Dazzi, P., Tolomei, G., Baraglia, R., Silvestri, F., and Orlando, S. *Challenges in designing an interest-based distributed aggregation of users in P2P systems*. In Proc. of IEEE ICUMT 2009, pp. 1–8.
- Tolomei, G. *Search the web x.0: mining and recommending web-mediated processes*. In Proc. of ACM RecSys 2009, pp. 417-420 [**rank = B**].
- Miori, V., Tarrini, L., Manca, M., and Tolomei, G . *DomoNet: a Framework and a Prototype for Interoperability of Domotic Middlewares based on XML and Web Services*. In Proc. of IEEE ICCE 2006, pp. 117–118.

TECHNOLOGY TRANSFER

Patent US20170154356A1 (Co-Author)
Title: “Generating actionable suggestions for improving user engagement with online advertisements”.
Details: https://patents.google.com/patent/US20170154356A1/en
Patent US20170004542A1 (Co-Author)
Title: “Method and system for providing content supply adjustment”.
Details: https://patents.google.com/patent/US20170004542A1/en
Patent US20170004541A1 (Co-Author)
Title: “Method and system for analyzing user behavior associated with web contents”.
Details: https://patents.google.com/patent/US20170004541A1/en
Patent US20180247222A1 (Co-Author)
Title: “Changing machine learning classification of digital content”.
Details: https://patents.google.com/patent/US20180247222A1/en

Spin-off at Università degli Studi di Padova (Co-founder)
Name: "CAPTCHAd"
Description: CAPTCHAd aims to allow entities which make use of CAPTCHA services (e.g., web portals, blogs, and – more generally – any service that needs to prevent automatic software bots to access their resources) to monetize by means of "sponsored challenges", namely CAPTCHA challenges that embed advertising contents.

TEACHING

Theory of Algorithms [48 hours] *2021 -*
M.Sc. in Applied Mathematics (1st year, 2nd semester)
Sapienza University of Rome, Italy

Big Data Computing [60 hours] *2020 - 2021*
M.Sc. in Computer Science (2nd year, 2nd semester)
Sapienza University of Rome, Italy

Operating Systems [60 hours] *2019 - 2020*
B.Sc. in Computer Science (2nd year, 1st semester)
Sapienza University of Rome, Italy

Python Programming for Data Science [40 hours] *2018 - 2019*
(within **Fundamentals of Information Systems**)
M.Sc. in Data Science (1st year, 1st semester)
Università degli Studi di Padova, Italy

Introduction to Computer Programming [32 hours] *2018 - 2019*
B.Sc. in Computer Science (1st year)
Università degli Studi di Padova, Italy

Python Programming for Data Science [44 hours] *2017 - 2018*
(within **Fundamentals of Information Systems**)
M.Sc. in Data Science (1st year, 1st semester)
Università degli Studi di Padova, Italy

Introduction to Computer Programming [32 hours] *2017 - 2018*
B.Sc. in Computer Science (1st year)
Università degli Studi di Padova, Italy

Database [teaching assistant - 25 hours] *2016 - 2017*
B.Sc. in Computer Science (2nd year)
Università degli Studi di Padova, Italy

Java Enterprise Edition [60 hours] *01/2014 - 02/2014*
Master "SIVE Formazione"
Università Ca' Foscari Venezia, Italy, at SIPE S.r.l.

OTHER EXPERIENCES

Software Engineer *07/2006 - 01/2008*

Company: Sysdat Informatica s.r.l., Pisa, Italy
Main responsibilities: Analysis, development, test, and deployment of Java Enterprise (J2EE-compatible) applications, specifically designed for third-party customers.

Projects:

- *Supermarket logistics*: Developed a web-based software to manage the dispatching of goods for the logistic department of "Conad" (i.e., one of the largest supermarket chains in Italy). Regularly interfacing with clients during the whole development stage up to the first product release.
- *Motorway payment system*: Developed the invoicing software system for "Autostrade per l'Italia S.p.A." (i.e., the Italian Concessionaire for toll motorway construction and management handling about 5 million daily customers, on average).

Software Engineer *12/2005 - 06/2006*

Company: NETikos S.p.A., Pisa, Italy
Main responsibilities: Analysis, development, test, and deployment of Java Enterprise (J2EE-compatible) applications, specifically designed for third-party customers.

Projects:

- *Mobile network provider web portal*: Developed the online recharge secure system for prepaid SIM cards of private customers inside the web portal of "Telecom Italia Mobile" (i.e., the largest Italian mobile telecommunications company counting more than 30 million subscribers). Implemented the electronic shopping cart and the interaction with the electronic payment gateway *GestPay*, powered by "Banca Sella S.p.A."

TECHNICAL SKILLS

Programming Languages	Python, Java, C/C++, R, Unix scripting (`bash`, `awk`, `sed`, etc.), SQL, Pig Latin, HiveQL, PHP, JavaScript
Libraries (Python)	Keras, Scikit-Learn, Pandas, Numpy, Scipy, Matplotlib, Seaborn
Development Environments	JupyterLab, Eclipse, NetBeans
Frameworks	Hadoop, Spark
Database	MySQL
Other Technologies	HTML/CSS, Hadoop, Git, SVN, LaTeX

ADMINISTRATIVE STRENGTHS

European Research Projects *01/2008 - 01/2012*

Name: FP7 Network of Exellence S-CUBE: Software Services and Systems Network
Unit: ISTI-CNR, Pisa, Italy
Description: Funded by the Commission of European Communities Information Society and Media Directorate-General, the project mission is to establish a unified, multidisciplinary, vibrant research community which will enable Europe to lead the software-services revolution and help shape the software service based Internet which will underpin the whole of our future society. S-Cube aims to push the frontiers of research in Service Oriented Computing by creating a vigorous research agenda where knowledge from diverse research communities is meaningfully synthesized, integrated and applied.
Activity: ISTI-CNR has been responsible for designing novel knowledge extraction techniques from service-oriented architecture system logs.
Details: http://www.s-cube-network.eu/

Event Organization

- Executive Director of the "2019 International Summer School on Machine Learning and Security"
  Details: 2019 ML&S School
- Keynote Speaker at the "Yahoo Tech Pulse 2016" Conference
  Title: "A Taste of Machine Learning"
- General Program Chair of the "IEEE Security and Privacy in Digital Advertising Workshop" (IEEE CNS SPA 2017)
  Details: IEEE CNS SPA 2017
- Program Co-Chair of the "IEEE Cyber-Physical Systems Security Workshop" (IEEE CNS CPS-Sec 2018)
  Details: IEEE CNS CPS-Sec 2018
- Program Co-Chair of the "IEEE Cyber-Physical Systems Security Workshop" (IEEE CNS CPS-Sec 2017)
  Details: IEEE CNS CPS-Sec 2017

Program Committee Membership

Conferences	ACM WSDM 2017-2021, ACM CIKM 2017-2019, ACM KDD 2019, ACM SIGIR 2018-2020, ACM ASONAM 2015-2020, ACM ICTIR 2017-2018, ECML/PKDD 2018-2019, TheWebConf (former WWW) 2018-2020, IJCAI 2016
Workshops	IIR 2015-2018