MANOS PAPAGELIS

YORK UNIVERSITY, LASSONDE SCHOOL OF ENGINEERING
ELECTRICAL ENGINEERING & COMPUTER SCIENCE (LAS 3050)
4700 KEELE STREET, M3J 1P3, TORONTO, ONTARIO, CANADA

WEB: HTTP://WWW.EECS.YORKU.CA/~PAPAGGEL
EMAIL: PAPAGGEL@EECS.YORKU.CA
TEL: (001) 416.736.2100 (EXT. 44782)

RESEARCH INTERESTS

My general research interests include:

A. DATA MINING
B. GRAPH MINING
C. TRAJECTORY DATA MINING
D. NATURAL LANGUAGE PROCESSING
E. DATABASES & KNOWLEDGE DISCOVERY
F. CITY SCIENCE / URBAN INFORMATICS

My research is primarily motivated by the need to fast and accurately discover patterns in large data sets that can inform data-driven decision making and data-intensive applications. It profoundly involves the design and development of methods at the intersection of **machine learning/AI**, **algorithms** and **data systems**. The rest of the document provides a bio sketch, a publication summary, a list of research-related awards and affiliations, an overview of my research work in each of these areas, and a justification for selection of publication venues.

BIO SKETCH – RESEARCH

I am currently an Assistant professor of Electrical Engineering and Computer Science (EECS) at York University, Toronto, Canada. My research interests include graph mining, data mining, big data and knowledge discovery, and city science / urban informatics. I hold a Ph.D. in Computer Science from the University of Toronto, Canada, and a M.Sc. and a B.Sc. in Computer Science from the University of Crete, Greece. Prior to joining York University, I was a postdoctoral fellow at the University of California, Berkeley. In the past, I have interned twice at Yahoo! Labs, Barcelona and worked as a research fellow at the Institute of Computer Science, FORTH, Greece. My research has appeared in top-tier journals in data mining, including the ACM Trans. on Knowledge Discovery from Data (ACM TKDD) and the IEEE Trans. on Knowledge and Data Engineering (IEEE TKDE), I have been granted one U.S. patent and have applied for two more. I have taught at the University of Crete, Greece, at the University of California, Berkeley, at the University of Toronto, and at York University.

PUBLICATIONS SUMMARY (AS OF SEP 2020)

Number of citations (by Google scholar):	1220+
H-index (by Google scholar):	13
Journal articles (peer-reviewed):	10
Conferences papers (peer-reviewed):	26
Workshop papers (peer-reviewed):	5
Magazine articles (peer-reviewed):	1
Patents (granted / applied for):	5 (1 / 4)
Citation analysis (by Google scholar):	http://goo.gl/3bi9m

RESEARCH-RELATED AWARDS

JUN 2020 21ST IEEE INTERN. CONF. ON MOBILE DATA MANAGEMENT (IEEE MDM 2020) – **Best Paper Award**
JUN 2018 19TH IEEE INTERN. CONF. ON MOBILE DATA MANAGEMENT (IEEE MDM 2018) – **Best Paper Award**
JAN 2018 26TH ACM INTERN. CONF. ON INFORM. & KNOWL. MANAGEMENT (ACM CIKM 2017) – **Outstanding Reviewer Award**
JAN 2011 ELSEVIER ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE (EAAI) – **Top Cited Article 2005-2010 Award**

RESEARCH-RELATED AFFILIATIONS

Member, EECS Department Faculty, Lassonde School of Engineering, York University
Member, Graduate Program in Electrical Engineering & Computer Science, York University
Member/PI, Data Mining Lab, EECS Department, York University
Member/Co-PI, BRAIN Alliance (Big Data Research, Analytics, Information Networks)
Member/Co-PI, Data Visualization and Analytics Training Program (NSERC CREATE DAV)
Member/Co-PI, Dependable Internet of Things Applications (NSERC CREATE DITA)
Member/Co-PI, Center for Innovation in Computing @ Lassonde (IC@L)

RESEARCH OVERVIEW[1]

A. DATA MINING

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Our research work in this area includes:

EFFICIENT MINING AND EXPLORATION OF MULTIPLE AXIS-ALIGNED INTERSECTING OBJECTS (2017–2020). Identifying and quantifying the size of multiple intersections among a large number of axis-aligned geometric objects is an essential computational geometry problem. The ability to solve this problem can effectively inform a number of spatial data mining methods and can provide support in decision making for a variety of applications. Currently, the state-of-the-art approach for addressing such intersection problems resorts to an algorithmic paradigm, collectively known as the sweep-line algorithm. However, its application on specific instances of the problem inherits a number of limitations. With that mind, we design and implement a novel, exact, fast and scalable yet versatile, sweep-line based algorithm, named SLIG. Our algorithm can be employed in a number of problems and applications involving the efficient computation of numerous axis-aligned object intersection problems in multiple dimensions. The key idea of our algorithm lies in constructing an auxiliary data structure when the sweep line algorithm is applied, an intersection graph. This graph can effectively be used to provide connectivity properties among overlapping objects, as well as to inform the much harder problem of finding the location and size of the common area defined by multiple overlapping objects. A thorough experimental evaluation on synthetic data of various characteristics and sizes, demonstrates that SLIG performs significantly faster than classic sweep-line based algorithms. SLIG is not only faster and more versatile, but also provides a suite of powerful querying capabilities. Results have been published in the 19th IEEE International Conference on Data Mining (ICDM 2019) [C21].

MRSWEEP: DISTRIBUTED IN-MEMORY SWEEP-LINE FOR SCALABLE OBJECT INTERSECTION PROBLEMS (2018–2020). Several data mining and machine learning problems can be reduced to the computational geometry problem of finding intersections of a set of geometric objects, such as intersections of line segments or rectangles/boxes. Currently, the state-of-the-art approach for addressing such intersection problems in Euclidean space is collectively known as the sweep-line or plane sweep algorithm and has been utilized in a variety of application domains, including databases, gaming and transportation, to name a few. The idea behind sweep line is to employ a conceptual line that is swept or moved across the plane, stopping at intersection points. However, to report all K intersections among any N objects, the standard sweep line algorithm (based on the Bentley-Ottmann algorithm) has a time complexity of $O((N + K)logN)$, therefore cannot scale to very large number of objects and cases where there are many intersections. In this paper, we propose MRSweep and MRSweep-D, two sophisticated and highly scalable algorithms for the parallelization of sweep-line and its variants. We provide algorithmic details of fully distributed in-memory versions of the proposed algorithms using the MapReduce programming paradigm in the Apache Spark cluster environment. A theoretical analysis of the proposed algorithms is presented, as well as a thorough experimental evaluation that provides evidence of the algorithms' scalability in varying levels of problem complexity. Results have been published in the 7th IEEE International Conference on Data Science and Advanced Analytics (IEEE DSAA 2020) [C25].

ELASTIC BULK SYNCHRONOUS PARALLEL MODEL FOR DISTRIBUTED DEEP LEARNING (2018–2020). The *bulk synchronous parallel* (BSP) is a celebrated *synchronization model* for general-purpose parallel computing that has successfully been employed for distributed training of machine learning models. A prevalent shortcoming of the BSP is that it requires workers to wait for the straggler at every iteration. To ameliorate this shortcoming of classic BSP, we propose ElasticBSP, a model that aims to relax its strict synchronization requirement. The proposed model offers more flexibility and adaptability during the training phase, without sacrificing on the accuracy of the trained model. We also propose an efficient method that materializes the model, named ZipLine. The algorithm can effectively balance the trade-off between quality of convergence and iteration throughput, in order to accommodate different environments or applications. A thorough experimental evaluation demonstrates that our proposed ElasticBSP model converges faster and to a higher accuracy than the classic BSP. It also achieves comparable (if not higher) accuracy than the other sensible synchronization models. Results have been published in the 19th IEEE International Conference on Data Mining (ICDM 2019) [C20]. An extended version is under review in a top-tier data mining journal [J8].

OL-HEATMAP: EFFECTIVE DENSITY VISUALIZATION OF MULTIPLE OVERLAPPING RECTANGLES (2018–2020). Visualization of the density of multiple overlapping axis-aligned objects is a challenging computational problem that can inform large-scale visual analytics, in diverse domains. For example, when dealing with crowd simulations, we care about constructing interaction maps, and in urban planning we care about city areas mostly frequented by people, to name a few. The primary objective of this research is, given a large set of axis-aligned objects, to devise efficient and effective data visualization methods that inform *whether*, *where* and *how much* these objects overlap. Currently, such visualizations rely on inefficient implementations of determining the size of the overlapping objects that do not scale well and are hard to accomplish. Approximate methods have also been proposed in the literature. To the contrary of these approaches, we aim to address this problem by exploiting state-of-the-art computational geometry methods based on the sweep line paradigm. These methods are fast and can determine the exact size of the overlap of multiple axis-aligned objects, therefore can effectively inform the visualization method. Towards

[1] All citations refer to the list of publications as they appear in the Curriculum Vitae (CV)

that end, we present OL-HeatMap, a novel type of a heat-map visualization that can be used to represent and perceive density of overlapping objects. Our experimental evaluation demonstrates the effectiveness of the proposed method in terms of both accuracy and running time, in a varying number of settings. Results of this work are currently under review at a top-tier big data analytics journal [J9].

SCENE CLASSIFICATION USING WORD REPRESENTATION LEARNING (2017–2018). Scene Classification has been addressed with numerous techniques in the computer vision literature. However, with the increasing number of scene classes in datasets in the field, it has become difficult to achieve high accuracy in the context of robotics. In this research, we approach the scene classification problem by bringing together traditional deep learning techniques and word representation learning methods in order to generate a word-embedding based scene classification algorithm. The key idea is that the context of an image (defined by objects detected in the scene) should be providing a strong signal of the correct scene label, and therefore can be employed to assist the scene class prediction problem. Objects present in the scene are represented by vectors and the images are re-classified based on the objects present in the scene to refine the initial classification obtained by a Convolutional Neural Network (CNN). We address indoor scene classification tasks using a model trained with a reduced pre-processed version of the Places365 dataset and an empirical analysis is done on a real-world dataset built by capturing image sequences using a GoPro camera. We also show a deployment of our approach on a robot operating in a real-world environment. Results have been published in the IEEE International Conference on Robotics and Automation - Multimodal Robot Perception Workshop (ICRA 2018 Workshops) [W4].

B. GRAPH MINING

Many data mining and machine learning problems can be formalized as graph problems. Our research work on graphs includes the following.

EVOLVING NETWORK REPRESENTATION LEARNING (2017–PRESENT). Large-scale network mining and analysis is key to revealing the underlying dynamics of networks, not easily observable before. Lately, there is a fast-growing interest in learning low-dimensional continuous representations of networks that can be utilized to perform highly accurate and scalable graph mining tasks. A family of these methods is based on performing random walks on a network to learn its structural features before feeding the sequence of random walks in a deep learning architecture to learn a network embedding. While these methods perform well, they can only operate on static networks. However, in real-world, networks are evolving, as nodes and edges are continuously added or deleted. As a result, any previously obtained network representation will now be outdated having an adverse effect on the accuracy of the data mining task at stake. The naive approach to address this problem is to re-apply the embedding method of choice every time there is an update to the network. But this approach has serious drawbacks. First, it is inefficient, because the embedding method itself is computationally expensive. Then, the data mining task results obtained by the subsequent network representations are not directly comparable to each other, due to the randomness involved in the new set of random walks involved each time. In this research, we proposed EvoNRL, a random-walk based method for learning representations of evolving networks. The key idea of our approach is to first obtain a set of random walks on the current state of network. Then, while changes occur in the evolving network's topology, to dynamically update the random walks in reserve, so they do not introduce any bias. That way we are in position of utilizing the updated set of random walks to continuously learn accurate mappings from the evolving network to a low-dimension network representation. Moreover, we present an analytical method for determining the right time to obtain a new representation of the evolving network that balances accuracy and time performance. A thorough experimental evaluation is performed that demonstrates the effectiveness of our method against sensible baselines and varying conditions. Results of this work have been published in the 7th International Conference on Complex Networks and Their Applications (Complex Networks 2018) [C19]. An extended version has been published in a special issue on "Machine Learning with graphs" of the Elsevier Applied Network Science (APNS) journal [J7].

MINING STREAMING AND DYNAMIC GRAPHS (2017–PRESENT). Large-scale graph mining and analysis is key to revealing the underlying dynamics of networks, not easily observable before. The conventional computational approach for performing graph/network analysis assumes there is a static network topology (and/or data) that is provided as input to a graph algorithm, which (always) terminates (i.e., produces an outcome or fails). Analyzing massive graphs via classical algorithms casts its own unique challenges (e.g., memory/time overhead), but the conventional approach is mostly insufficient for many modern data processing needs. Over the last years, there has been considerable interest in designing algorithms for processing graphs in the data stream model, where the input is defined by a stream of graph data (e.g., a stream of edges), and the graph algorithm, aware of these changes, must be able to accept the changes faster than a naive re-computation of an algorithm on static graphs. Algorithms in this model must operate under specific constraints: (i) the input stream must be processed in the order it arrives, and (ii) the processing can only use a limited amount of memory. This paragraph describes work in progress.

GRAPH AUGMENTATION (2011–2015). During my second research internship at Yahoo! Research, Barcelona, I became interested to the problem of suggesting a number of non-existing edges to add in a graph in order to optimize a connectivity problem. It is known that slight modifications in the network topology of a graph, might have a dramatic effect on its connectivity and thus to its capacity to carry on social processes, such as information cascades. We approached network modification as a graph augmentation problem where we seek to find a few non-existing edges to add to the graph. To this end, we presented methods that can accurately and efficiently evaluate the importance of non-existing edges to guide the graph augmentation process. In the augmented graph (i) social processes evolve faster and can reach more nodes, and (ii) random walks can converge a multitude times faster, giving rise to faster graph simulations. Preliminary results of this work

have been published at the 20th ACM International Conference on Information and Knowledge Management (ACM CIKM 2011) conference [C10]. More recent research results have appeared in the Hellenic Data Management Symposium (HDMS) [C12] and have been published in the ACM Transactions on Knowledge and Discovery from Data journal (ACM TKDD) [J4].

SOCIAL SEARCH (2008–2013). The main idea of social search is to utilize social cues available in online social networks to improve search results, towards personalized search. An interesting, but challenging data mining task in (static and dynamic) social networks is given a specific user (or node) to probe the nodes (and/or the information they are associated with) that are found a few hops away from that user. We presented sampling-based algorithms that given a user in a social network can quickly obtain a near-uniform random sample of users in its neighborhood. Further, we employed these algorithms to quickly approximate the number of users in a user's neighborhood that have endorsed an item. Our methods are highly accurate and very efficient and can be utilized in a number of applications where we aim to rank items in a network, such as, to improve the quality of the user's search experience in a social search engine, or to improve the accuracy of collaborative filtering methods in a recommendation engine. Preliminary results were published at the ACM CIKM/SSM Workshop (Search in Social Media) [W3]. More recent results have been published in the IEEE Transactions on Knowledge and Data Engineering journal [J3].

SOCIAL INFLUENCE AND DIFFUSION IN ONLINE SOCIAL MEDIA (2009–2011). In the presence of social influence, an idea, behavior norm, or product diffuses through the social network like an epidemic. Analysis of massive data sets at an aggregate level raises interesting observations of information cascades within social systems, but it has its limitations as it typically does not allow for interpretations at the individual level. More importantly, it does not allow to argue about the causality of the observed cascades. During my first internship at Yahoo! Labs, my research was driven by more refined research questions related to information cascades. Why cascades happen? How to determine whether cascades will occur in a social system? How one can influence a cascade process? Social scientists agree that in the presence of social influence, an idea or behavior diffuses through a social network like an epidemic. In our research, we were interested to develop methods that given a history of individual actions and social relationships of individuals can qualitatively and quantitatively detect social influence effects in a social system. Further, we developed a basic framework that allows to study the causality between individual actions of users and the social influence that they exert to their social network. As a case study for our research we employed a popular online social graph, Flickr (including 525K nodes, 47M edges), and our analysis of social influence was based on the cascade of the geotagging behavior. To the best of our knowledge, our work was one of the first that tried to bridge the gap between aggregate level cascade processes and individual behavior using large data sets. Results of this work have been published at the ACM Hypertext 2011 conference [C9], a U.S. patent has been granted [P1] and another one is under review [P2].

C. TRAJECTORY DATA MINING

Mining large-scale trajectory data streams (of moving objects) has attracted significant attention due to an abundance of modern tracking devices and a number of real-world applications. Our research work in this area includes:

MINING OF NODE IMPORTANCE IN TRAJECTORY NETWORKS (2017–2019): In this research, we are interested in evaluating the relative importance of such objects through monitoring their interactions with other objects, over time. Which object has encountered more other objects? When did these encounters happen and how long they lasted for? To address this type of questions, we consider a trajectory network that is defined based on the proximity of moving objects over time. Given a trajectory network we are able to evaluate the network importance of an object by monitoring its complex network connections to other nodes over time. Traditional approaches to address the problem heavily rely on evaluating network metrics over a number of static network snapshots, streaming algorithms that focus on simple network metrics or expensive trajectory similarity and clustering methods that may need further post-processing. In contrast to these approaches, we devise a method that is able to simultaneously evaluate a number of network metrics of interest for all moving objects (i.e., all trajectories), over time. Our proposed method is based on efficiently computing and representing the interactions of objects over a period of time, under varying conditions. Then, a fast and accurate one-pass algorithm is used to extract the metrics of interest, all at once. Through experiments on various types of synthetic data, we demonstrate the effectiveness of our methods against sensible baselines, for a varying range of conditions. Our proposed methods are easy to implement and adapt in different scenarios and domain-specific applications. Results of this work have been published in the 6th IEEE International Conference on Big Data (IEEE BigData 2018) [C18]. Extensions of this work with application in epidemic spreading modeling are currently under review at a network analytics journal [J10].

GROUP PATTERN DISCOVERY OF PEDESTRIAN TRAJECTORIES (2017–2019): In this research, we are interested in mining group patterns of moving objects. Group pattern mining describes a special type of trajectory mining task that requires to efficiently discover trajectories of objects that are found in close proximity to each other for a period of time. In particular, we focus on trajectories of pedestrians coming from motion video analysis and we are interested in interactive analysis and exploration of group dynamics, including various definitions of group gathering and dispersion. Towards this end, we presented a suite of (three) tensor-based methods for efficient discovery of evolving groups of pedestrians. Traditional approaches to solve the problem heavily rely on well-defined clustering algorithms to discover groups of pedestrians at each time point, and then post-process these groups to discover groups that satisfy specific group pattern semantics, including time constraints. In contrast, our proposed methods are based on efficiently discovering pairs of pedestrians that move together over time, under varying conditions. Pairs of pedestrians are subsequently used as a building block for effectively discovering groups of pedestrians.

The suite of proposed methods provides the ability to adapt to many different scenarios and application requirements. Furthermore, a query-based search method is provided that allows for interactive exploration and analysis of group dynamics over time and space. Through experiments on real data, we demonstrated the effectiveness of our methods on discovering group patterns of pedestrian trajectories against sensible baselines, for a varying range of conditions. In addition, a visual testing is performed on real motion video to assert the group dynamics discovered by each method. Results of this work have been published in the 19th IEEE International Conference on Mobile Data Management (IEEE MDM 2018) [C16, C15], where it is awarded **the best paper award**. An extension of this work has been published in a special issue of the GeoInformatica journal [J6].

Learning Semantic Relationships of Geographical Areas based on Trajectories (2018–2020): In this research, we focus on leveraging the abundance of trajectory data to automatically and accurately learn latent semantic relationships between different geographical areas (e.g., semantically correlated neighborhoods of a city) as revealed by patterns of moving objects over time. While previous studies have utilized trajectories for this type of analysis at the level of a single geographical area, the results cannot be easily generalized to inform comparative analysis of different geographical areas. In this work, we study this problem systematically. First, we present a method that utilizes trajectories to learn low-dimensional representations of geographical areas in an embedded space. Then, we develop a statistical method that allows to quantify the degree to which real trajectories deviate from a theoretical null model. The method allows (a) to distinguish *geographical proximity* to *semantic proximity*, and (b) to inform a comparative analysis of two (or more) models obtained by trajectories defined on different geographical areas. This deep analysis can improve our understanding of how space is perceived by individuals and inform better decisions of urban planning. Our experimental evaluation demonstrated the effectiveness and usefulness of the proposed statistical method in two large-scale real-world data sets coming from the New York City and the city of Porto, Portugal, respectively. The methods we present are generic and can be utilized to inform a number of useful applications, ranging from location-based services, such as point-of-interest recommendations, to finding semantic relationships between different cities. Results of this work have been published in the 21st IEEE International Conference on Mobile Data Management (IEEE MDM 2020) [C22], where it is awarded **the best paper award**.

D. Natural Language Processing

Our research work in this area focuses on sentiment analysis, which aims to extract subjective information about the polarity of a set of documents, such as online reviews. Our research focus on emotion detection that can inform affective tasks and includes the following.

A Comprehensive Analysis of Preprocessing for Word Representation Learning in Affective Tasks (2019–2020). Affective tasks such as sentiment analysis, emotion classification and sarcasm detection have been popular in recent years due to abundance of user-generated data, accurate computational linguistic models, and broad range of relevant applications in various domains. At the same time, many studies have highlighted the importance of text preprocessing, as an integral step to any natural language processing prediction model and downstream task. While preprocessing in affective systems is well-studied, preprocessing in word vector-based models applied to affective systems, is not. To address this limitation, we conduct a comprehensive analysis of the role of preprocessing techniques in affective analysis based on word vector models. Our analysis is the first of its kind and provides useful insights of the importance of each preprocessing technique when applied at the training phase, commonly ignored in pretrained word vector models, and/or at the downstream task phase. Results have been published in the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020) [C23].

Emotion-enriched Word Representation Learning (2017–2019). Most word representation methods consider the lexical semantics and syntax based on the co-occurrences of words. As a consequence, emotionally dissimilar words such as "happy" and "sad" occurring in similar contexts are estimated to be more similar than the emotionally similar word pair "happy" and "joy", which leads to rather undesirable consequences in affective tasks, such as emotion classification. In order to address this limitation, we propose a novel method of obtaining emotion-enriched word representations, which projects emotionally similar words into neighboring spaces. The proposed approach leverages distant supervision to automatically obtain a large training dataset of text documents and two recurrent neural network architectures for learning the emotion-enriched representations. In extensive evaluation on two tasks including emotion classification and emotion similarity, the proposed representations outperform several competitive generic as well as affective representations. Results have been published in the 27th International Conference on Computational Linguistics (COLING 2018) [C17]. We have also worked on the harder problem of sarcasm detection in text and results have been published in the 28th International Conf. on Computational Linguistics (COLING) [C26] and in the 43rd ACM Intern. Conf. on Research and Development in Information Retrieval (ACM SIGIR 2020) [C24].

Leveraging Emotion Features in News Recommendations (2017–2019). Online news reading has become very popular as the web provides access to news articles from millions of sources around the world. As a specific application domain, news recommender systems aim to give the most relevant news article recommendations to users according to their personal interests and preferences. Recently, a family of models has emerged that aims to improve recommendations by adapting to the contextual situation of users. These models provide the premise of being more accurate as they are tailored to satisfy the continuously changing needs of users. However, little attention has been paid to the emotional context and its potential on improving the accuracy of news recommendations. The main objective of this research is to investigate whether, how and to what extent emotion features can improve recommendations. Towards that end, we derive a large number

of emotion features that can be attributed to both items and users in the domain of news. Then, we devise state-of-the-art emotion-aware recommendation models by systematically leveraging these features. We conducted a thorough experimental evaluation on a real dataset coming from news domain. Our results demonstrate that the proposed models outperform state-of-the-art non-emotion-based recommendation models. Our study provides evidence of the usefulness of the emotion features at large, as well as the feasibility of our approach on incorporating them to existing models to improve recommendations. Results have been published in the 7th International Workshop on News Recommendation and Analytics. (INRA / ACM Recommender Systems 2019 Workshops) [W5].

E. DATABASES AND KNOWLEDGE DISCOVERY

ANALYSIS OF PATTERNS OF COMMUNICATION IN SOCIAL MEDIA (2007–2010). With an increasing number of people that read, write and comment on blogs, the blogosphere has established itself as a dynamic medium of communication. In our research we are investigating patterns of communication in social media. In particular, we have conducted a thorough analysis of how information cascades in the blogosphere (i.e., the succession of linking behavior through which information propagates among blogs). Our analysis utilized one of the largest data sets available at the time (including 30M blogs and over 400M posts) and was distilled into a comprehensive report that contributed to a better understanding of the overall linking activity in the blogosphere. In particular, we analyzed: (i) trends on the degree of engagement and reaction of bloggers in stories that become available in blogs, (ii) the structure of cascades that are attributed to different population groups constrained by factors of gender, age, and continent, and (iii) topic-sensitive information cascades. Our analysis was the first to incorporate heterogeneity (differences in the groups) and revealed notable variations in the structural properties of cascades and in the ability of blogs to trigger prompt and widely-spread cascades that mainly depend on the blogger's profile and the topic of a post. Analyzing information cascades can be useful in various domains, such as providing insight of public opinion and developing better prediction models with applications in health, business, politics and more. More importantly, our study established evidence of a cascading effect of online communication occurring in online social media and stimulated the next research questions. Results of this work were published at the 3rd Int. AAAI Conf. on Weblogs and Social Media (ICWSM 2009) [C8].

ONLINE SOCIAL NAVIGATION (2007–2008). The primitive way to access all online information remains the web browser. But, web browsing or navigation is commonly assumed to be an autonomous and passive process where a user interacts with information but not with other people. Motivated by this fact we focused on ways to make web browsing more social. In our research, we considered a "social navigation" paradigm for browsing the web. Social navigation is the process where a number of people that share interests and searching goals decide to coordinate their efforts. With that in mind, we designed and developed a practical social navigation system that makes users aware of other users accessing similar information at similar time and encourages interpersonal real-time communication and collaboration. Preliminary results of this work were published at the ACM Hypertext Conference [C6]. More recent results were published at the IEEE/WIC/ACM Web Intelligence Conference [C7].

RECOMMENDATION ALGORITHMS (2003–2005). Study of recommendation algorithms has been a long-term agenda item that has led to a number of publications and to the development of a recommendation system, which served both as a research platform and as a free online service. First, we investigated the way in which recommendation algorithms can be employed in order to discover dynamic, virtual, online communities [C1]. We spent the next period developing and evaluating the quality of collaborative filtering recommendation algorithms that are based on item similarities, instead of user similarities. The results appeared in the CIA Workshop [W1], HDMS 2004 [C2] and, after invitation, to a special issue of the Elsevier Engineering Applications of Artificial Intelligence Journal [J1]. Next, we turned focus on the scalability problem of recommendation algorithms. We developed a method of incremental computation of user similarities that was improving the performance of recommendation algorithms without reducing their quality. Results were published in the proceedings of the International Symposium on Methodologies for Intelligent Systems (ISMIS) [C3]. Then, we investigated trust implications in web based social networks. We were particularly interested in modeling trust, developing a computational model for trust, and investigating the way in which trust fits architectonically into large scale information discovery systems that function within highly-distributed environments. We argued that trust propagation techniques could be efficiently employed so as to alleviate the sparsity problem of recommendation systems. Results of this work were published at the International Conference on Trust Management (iTrust) [C4].

INFORMATION INTEGRATION (2003–2007). While at ICS-FORTH, I was engaged in the "Ubi-Erat-Lupa" EU funded research project. The objectives of the project were to interlink archaeological research on the Roman era systematically and transnationally and to exchange expert knowledge. From a computer scientist's viewpoint the project engaged many information integration challenges. I worked with Dr. Martin Doerr towards the automatic place name identification problem, a complex problem when integrating geospatial information. We developed a method that permits to estimate the precision of place name matching, the completeness of a gazetteer and the semantic inconsistency that relies in a digital gazetteer. Preliminary results have been presented in the NKOS Workshop of the European Conference on Digital Libraries [W2]. Complete research results appeared in the IEEE Trans. on Knowl. and Data Engin. Journal (TKDE) [J2]. In the scope of this project, I had the opportunity to become familiar with technologies for semantic web information integration and querying.

COMPUTER SUPPORTED COLLABORATIVE WORK (2003–2006). Besides my interest in research, I am very fascinated in building large-scale information systems with research extensions. Among others, such as the Movie Recommendation System dedicated to research purposes and an Online Questionnaire System, I would like to especially refer to "Confious". Confious is a state-of-the-art conference management system that combines modern design, sophisticated algorithms and powerful engine so as to efficiently and professionally support the whole submission and reviewing process of an academic/research conference or workshop. Confious was initially selected by the University of Crete, Greece, as a student's innovative idea worth funding for further development. An industrial paper describing Confious was published at the WISE Conference [C5], while it was also mentioned in the ERCIM News [M1]. Throughout the years, Confious has been successfully used to support the submission and reviewing process of a number of international conferences and workshops.

F. CITY SCIENCE / URBAN INFORMATICS

ONLINE SOCIO-TECHNICAL ANALYSIS OF SUSTAINABLE BUILDINGS (2015–2017). Professionals and researchers of the Architectural, Engineering & Construction (AEC) industry, as well as, public policy makers are challenged by the increasing complexity and need to improve our understanding of the social, technical and business dimensions of green building projects. This typically requires close cooperation of the design team, the architects, the engineers, and the rest of the stake-holders at all project stages, but most importantly availability of new methods, tools and strategies that are enabled by emerging technologies. This research builds around an online platform (Green2.0) that tries to leverage advancements in Building Information Models (BIM), energy-efficiency simulation tools and online social network analysis methods to enable a data-driven approach to building design, planning, construction and maintenance. The platform advances the current state of the art by providing an online integrated environment for (a) efficient storage, indexing, querying, 3D visualization and exploration of BIMs, (b) sharing BIMs and enabling online collaboration among the various stakeholders (c) interactive energy efficiency analysis of buildings by automatically linking IFC to external energy simulation libraries, and (d) interactive analysis of patterns of social interactions and collaboration networks of AEC professionals. A number of papers have been published that describe technical contributions of the system. Preliminary results were published at the International Conference on Civil, Structural and Transportation Engineering (ICCSTE 2015) [C11] and the International Conference on Smart Infrastructure and Construction (ICSIC 2016) [C14]. Extensions of this work have been published as a demo paper [C13], a full research paper to a top-tier AEC journal [J5].

SOFTWARE RELEASES

DATE	NAME	BRIEF DESCRIPTION	URL
2020	MRSWEEP	A distributed in-memory version of the popular sweep-line algorithm for finding axis-aligned geometric object intersections (overlaps). Implemented using MapReduce in Apache Spark.	https://github.com/tipech/mrsweep
2020	OL-HEATMAP	A novel heatmap-like visualization method for effective density visualization of the overlaps of multiple intersecting axis-aligned objects (line-segments in 1*D*, rectangles in 2*D*, cuboids in 3*D*, etc.).	https://github.com/tipech/overlapGraph
2019	ZIPLINE	An optimization method for materializing ElasticBSP. ElasticBSP is an efficient method for training deep learning models in a distributed environment.	https://github.com/xingzhaoo/ElasticBSP (currently confidential due to an NDA)
2019	SLIG	An efficient method for finding information about the overlaps of multiple intersecting axis-aligned objects (line-segments in 1*D*, rectangles in 2*D*, cuboids in 3*D*, etc.).	https://github.com/tipech/overlapGraph
2019	EVONRL	A deep learning method for obtaining continuous low-rank representations of an evolving graph/network.	https://github.com/farzana0/EvoNRL
2018	SLOT	An efficient method for profiling nodes in trajectory networks (dynamic networks defined by trajectories). Metrics include node degree, triangle membership and connected components – defined as a function of time.	https://github.com/tipech/trajectory-networks

SUMMARY OF CONTRIBUTIONS TO JOINT PUBLICATIONS

The rationale used for the order of authors and a summary of contributions to joint publications is provided below.

[J6], [J7], [J9], [J10], [C15], [C16], [C18], [C19], [C21], [C22], [C23], [C25], [C26], [W4], [W5]: These are publications co-authored with current and graduated students; it also includes joint publications with students in my class working on projects proposed and supervised by myself. As a junior professor, I adhere to close supervision, so I assumed active role in all stages of the research including problem ideas, problem formulation & analysis, methodology, algorithm design, experimental evaluation design, paper writing. Students were the principal contributors in algorithm development, running experiments, and knowledge dissemination. They also contributed in exchange of ideas and design of solutions, as well as, paper writing. By tradition in CS, students are listed as first authors (with the author list following degree of contribution) and the PI is listed last.

[C17], [C24]: This is a collaboration with Prof. Aijun An (YorkU) and one of her MSc students. The author list denotes degree of contribution.

[J8], [C20]: This is a collaboration with Prof. Aijun An (YorkU) and one of her MSc students. The author list denotes degree of contribution.

[J5], [C11], [C13], [C14], [P3]: These publications are related to a research project at the University of Toronto. In [J5], [C14] and [P3] authors are listed in an alphabetical order denoting equal contribution. The first author is the PI of the project and the second is a collaborator at the Technical University of Eindhoven. In [C11] and [C13] the order of the authors denotes degree of contribution.

[J3], [J4], [C8], [C9], [C10], [C12], [W3], [P1], [P2]: These publications are directly related to my PhD thesis. I assumed the principal role in all stages of the research (idea, problem formulation & analysis, methodology, algorithm design, algorithm development, experimental evaluation, paper writing and knowledge dissemination). My PhD supervisor at the University of Toronto also contributed in exchange of research ideas, problem formulation and proof-reading parts of the paper. In [J3] and [W3], the coauthors include a research collaborator at the University of Texas, Austin. Co-authors of [C9] and [C10] are my research supervisors while I was an intern at Yahoo! Labs; they also coordinated efforts of the patent applications [P1] and [P2].

[C6], [C7]: This is a collaboration with colleagues from the University of Patras. The order of the authors denotes degree of contribution.

[J2], [W2]: These publications are directly related to my research fellowship at ICS-FORTH, Greece. I assumed the principal role in all stages of the research (idea, problem formulation & analysis, methodology, algorithm design, algorithm development, experimental evaluation, paper writing and knowledge dissemination). My supervisor at ICS-FORTH also contributed in exchange of research ideas, problem formulation, writing and proof-reading parts of the paper.

[J1], [C1], [C2], [C3], [C4], [C5], [W1]: These publications are directly related to my MSc thesis. I assumed the principal role in all stages of the research (idea, problem formulation & analysis, methodology, algorithm design, algorithm development, experimental evaluation, paper writing and knowledge dissemination). My MSc supervisor at the University of Crete, Greece also contributed in exchange of research ideas, problem formulation and proof-reading parts of the paper. In [C3], [C4] and [C5] undergraduate students were part of the collaboration and were listed in the list of coauthors.

JUSTIFICATION FOR SELECTION OF PUBLICATION VENUES

Our research has been published at **the top scientific journals** in the area of **data mining and knowledge discovery**, including the IEEE Transactions in Knowledge and Data Engineering (**IEEE TKDE**) and the ACM Transactions of Knowledge Discovery from Data (**ACM TKDD**). It has also appeared in the Elsevier Engineering Applications of Artificial Intelligence (**EAAI**) journal, the journal of **Applied Network Science**, and **GeoInformatica**, all of which are **highly reputable journals of their discipline**. Our interdisciplinary research on mining social interactions around building information models (BIM) has been published at **the top journal** pertaining to the use of Information Technologies in Civil Engineering, **Automation in Construction**.

Our research has also been disseminated in **premier research conferences** on **information and knowledge management**, **natural language processing**, **databases and systems**, such as the IEEE International Conference on Data Mining (**IEEE ICDM**), the Annual Meeting of the Association for Computational Linguistics (**ACL**), the IEEE International Conference on Mobile Data Management (**IEEE MDM**), the ACM SIGIR Conference on Research and Development in Information Retrieval (**ACM SIGIR**), International Conference on Computational Linguistics (**COLING**), ACM International Conference on Information and Knowledge Management (**ACM CIKM**), ACM Conference on Hypertext and Social Media (ACM Hypertext), ACM Computer-Supported Cooperative Work (**ACM CSCW**), Web Information Systems Engineering (**WISE**), AAAI International Conference on Weblogs and Social Media (**ICWSM**) series of conferences.