UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

DIPARTIMENTO DI
INFORMATICA

PhD Program in Computer Science and Mathematics
XXXIII cycle

Research Project

PhD Student: Emanuele Pio Barracchia

Supervisor: Prof. Michelangelo Ceci

Coordinator: Prof. Maria F. Costabile

PhD student signature ______________________________

Supervisor signature ______________________________

1. Research title:

Predictive tasks in large dynamic networks

2. Research area:

Machine Learning and Data Mining: Big Data Analytics

3. Research motivation and objectives

The advent of social networks provided users with a new powerful mean to communicate and to express their ideas and preferences. All this data can be useful for industry in order to, for example, make new products consistent with user preferences. Data from social networks can be integrated with information obtained from different data sources, such as online newspaper and online blogs, in order to detect or predict more extended phenomena like rare disease outbreaks.

Such networks can be analyzed by methods in the Dynamic Network Analysis (or DNA) field, that brings together other traditional fields such as social network analysis, link analysis, social simulation and multi-agent systems within network science and network analysis.

Dynamic Network Analysis analyzes meta-networks, that are:

- *Multi-mode:* there are many types of nodes. In the case of a social network we can distinguish people, locations, organizations, etc.;
- *Multi-link:* there are many types of links between nodes. Examples from social networks are friendship, membership to a group, etc.;
- *Multi-level:* some nodes may be members of others nodes. For example, a node representing a person can have a link of type "membership to" to an organization where the person works.

In many cases, links in networks analyzed by DNA are not binary, but they are associated with a probability that they exist.

Networks like social networks are characterized by their dynamicity: their structure can change over and over in terms of insertion of new nodes, new links, new groups of nodes or deletion of old ones. In the case of social networks, the possible network structure changes can be addressed to new user registration, addition of a user to a group, review posted by user regard a shop, a pub or other places, etc.

In this project, the focus will be on the predictive tasks on such large and dynamic networks, addressed by means of Big Data Analytics techniques. Predictive tasks aim to exploit some independent variables in order to predict unknown or future values of other dependent variables, which can be categorical, numerical, or even complex structures (hierarchies, graphs, etc.).

The main goal of the research will be the development of methods able to predict the evolution of dynamic networks, in terms of future nodes or future links, by exploiting Dynamic Network Analysis techniques. In this project, different challenges of such a field will be investigated: 1) the effective and efficient analysis of large collections of network data; 2) the integration of data coming from unrelated and possibly heterogeneous sources of information; 3) the understanding of the possible effects of the interactions between two nodes on the other nodes of the network; 4) the development of general purpose predictive algorithms, which can be adopted in the context of different domains; 5) the development of algorithms able to adapt the identified networks over time after the arrival of new data; 6) the study of algorithms able to track groups of nodes or links in networks over time; 7) the forecasting of future changes in the networks.

4. State of the art

Dynamic networks are a type of networks whose structure, in terms of nodes and links, changes as a function of time. Examples of dynamic, or evolving, networks are social networks, like Facebook and Twitter, professional networks, like LinkedIn, transportation networks and telecommunication networks. In recent years, many works have been proposed to study the evolution of this type of networks, trying to predict future behaviors.

For example, the recent growth of social networks raised new challenges and opportunities in the spam detection landscape. In [1], Fakhraei et al. developed methods to identify spammers in evolving multi-relational social networks. To achieve their goal, the authors proposed a content-independent framework which can be used on (partially or completely) anonymized data. The framework consists of three phases: *i)* extraction of a directed graph for each type of relationship; *ii)* exploitation of the activity sequence of each user across these relationships; *iii)* use of a statistical relational model based on hinge-loss Markov random fields to perform collective reasoning using signals from an abuse reporting system in the social network and assigning credibility scores to the users who offer feedback via the reporting system. A possible extension of this work, proposed by authors, is to develop an online learning method that can incorporate the changing in the dynamic network without retrain it with every new sample. Another extensions would be classify or cluster spammer accounts based on their target accounts in order to improve the results. These extensions are motivated by the existence of spam campaigns.

In 2016, Muthiah et al. published a paper [2] that describes a system able to detect future planned protests. The system, called EMBERS (Early Model-Based Event Recognition using Surrogates), aims at detecting references to future planned events starting from the analysis of relevant news and social media, like Facebook and Twitter. Differently from other systems, EMBERS is able to extract dates also when they are expressed with terms like “tomorrow” or “next Monday”. Moreover, the proposed system can predict where the protest will be hold at city-level granularity.

In [3], Rames et al. presented a new framework able to create a model that represents and combines different kinds of user information extracted by a professional network. Authors, during the creation of the model, take into account different types of actions performed by users, such as moving jobs, adding a new skill, following content and adding oneself to groups. To achieve their goal, Rames et al. propose a method which exploits two different structures: *i)* a graph G = (V,E), where nodes in V represent users and edges in E are time-stamped edges between pair of users; *ii)* an action propagation graph for each action type. The action propagation graph is used to capture how users react to actions performed by their friends. In order to develop the model, Rames et al. used hinge-loss Markov Random Fields (HL-MRFs). HL-MRFs is able to combine different heterogeneous relationships (in this case one relationship type correspond to one action type) between individuals to learn pair-wise influence probabilities.

Another work is [4] by Rekatsinas et al. The goal of this paper is to develop techniques that can forecast the emergence and the progression of rare infectious diseases by combining data from different data sources, such as news articles, blogs, search engine logs, micro-blogging services, etc. To achieve this goal, authors introduce a new framework called SourceSeer. This framework is able to combine spatiotemporal topic models with source-based anomaly detection techniques. SourceSeer consists of two main tasks: *i)* analysis of past data to detect the spatiotemporal patterns of diseases; *ii)* prediction of future disease outbreaks. In the end, SourceSeer is able to combine predictions produced from different sources in order to compute a final prediction.

Social media services can be considered as the primary channel used by both individuals and organizations to communicate ideas, share opinions, promote resources and report on events. In [5]

Kumar et al. proposed a framework able to analyze microblogging data from Twitter, focusing on the political domain, aiming at the identification of political relationships. The proposed framework is able to create different unsupervised probabilistic models to infer strategic relationships between organizations in political domain, expressing strength of relations, incorporating social context and analyzing issues to infer relationships. A possible extension of this work, proposed by authors, would be to model the change in relationships over time.

In the end, another application of the analysis of dynamic networks is the survival analysis whose goal is to model the length of time until a particular event occurs. One of the tasks in survival analysis is to cluster people in a semi-supervised fashion by using their attributes and their survival times. This clustering procedure has been adopted in the literature in order to identify cancer subtypes from gene expression data. The survival analysis can also be applied in social network in order to study the user survivability, that is the time the user will stay in the system. In a work of Mouli et al. [6], the authors proposed a decision tree based algorithm that identifies clusters with significantly different distribution. The proposed method consists in the construction of a decision tree where every decision node is the attribute-value pair with the lowest p-value. In this way, after each split, the survival distributions are most likely different from each other. After the construction of the decision tree, the proposed algorithm clusters the leaf nodes in order to put in the same cluster those leaves with similar survival distributions.

5. Problem approach

The approaches followed in the project will try to find a solution to the challenges mentioned. In order to reach this goal, different techniques will be investigated:

- Hierarchical multi-type clustering in order to cluster heterogeneous nodes similar to each other in the same cluster. A hierarchy of heterogeneous clusters helps human experts to understand results. With the arrival of new data, such clusters could change in order to reflect the evolution of the network. In this way a possible implementation would be an algorithm that performs clustering every time new data arrives, using the centroids of previous clusters as labelled samples in order to keep memory of what learnt until then. Clusters have to reflect two different types of evolution:
    - *Topic evolution*: in this case clusters have to change in order to include a different number of instances;
    - *New topic discovery*: in this case the algorithm has to be able to discovery new topics identifying new rising clusters that are very dissimilar from that already determined;
- Hinge-loss Markov random fields [7], a new kind of probabilistic graphical model that generalizes different approaches to convex inference. To define Hinge-loss Markov random fields it is possible use Probabilistic Soft Logic (or PSL), a probabilistic language programming with a syntax based on first-order logic;
- Incremental algorithms in order to take into account the arrival of new data (new nodes or new edges) in the dynamic network. When new data is available, it could implicate the insertion or the deletion of nodes or edges in the network. However, every time new information is available, it could be interesting analyze how such new data impact on the network structure. For example, a possible analysis could concern the investigation of the potential changes on the weight associated to the edges or on the attributes of some nodes;
- Apache Spark, an open-source cluster-computing framework maintained by Apache Software Foundation. Apache Spark extends MapReduce giving the possibility to perform more types of computation, such as interactive queries and stream processing. Spark can be very useful with large networks because it increase its computational speed executing data processing in memory.

Moreover, to develop methods that solve link prediction tasks on data from real contexts, we plan to undertake the following steps:

1) Gather data from real contexts from different data source, such as newspapers, social networks, etc.;
2) Integrate such data;
3) Organize data in a network structure;
4) Implement algorithms which perform predictive tasks on the network. In particular, it would be interesting to predict possible insertions or deletions of edges and nodes, the changes of the weights associated to the edges, the creation or the breaking up of groups of nodes, etc.;
5) Manage and possibly exploit the evolution of the dynamic network developing incremental methods.

6. Expected results

The analysis of dynamic networks can be applied to different domains, such as:

- Bioinformatics, to study relationships between molecular entities or between molecular entities and diseases. For example, it can be interesting to study the correlation between non-coding RNAs and diseases in order to identify associations previously unknown.
  Such a study can be very useful for biologists because the investigation of all possible associations would entail expensive experiments in laboratory. However, this type of analysis can only help biologists to perform a more accurate research, but the task of determining which relationships is real remains their responsibility;
- Social network analysis, to identify spam users, possible customers of a new product, influencers to contact in order to promote a new product, to discover communities and to investigate on their evolution, to analyze user behaviors, to investigate on phenomena such as terrorism, bullying and mobbing, etc.;
- Geospatial analysis, to detect environmental changes, to perform analysis in transportation and epidemiology domains, etc.;
- Survival analysis, to predict the user survivability in the system or to identify disease subtypes starting from gene expression data or the amount of time the user will spend using a system;
- Politics, to detect, for example, civil unrest or to analyze how constituents are approaching to political factions.

In the end, the analysis of dynamic networks can be applied to both research domain and industrial domain.

In this project all the mentioned domains will be taken into account during the experimental evaluation in order to tailor algorithms, possible heuristics and evaluation measures for each specific application domain.

7. Phases of the project

1st year: study of the literature, of the state of art and other basic research material

Activity 1.1 study of the state of the art regarding Big Data Analytics with particular attention to models to represent Big data;

Activity 1.2 in-depth analysis concerning heterogeneous networks and dynamic networks;

Activity 1.3 attendance of courses and seminars included in the doctorate study plan;

Activity 1.4 participation to international schools and conferences regarding topics relevant for the research theme and the goals planned.

2nd year: development of methods

Activity 2.1 study of the works produced by other researchers with the same goals;

Activity 2.2 design and development of methods able to solve predictive tasks in dynamic networks;

Activity 2.3 evaluations of the developed methods, comparison with existing approaches and publication of the results.

3rd year: application to some domains and writing of the doctorate thesis

Activity 3.1 comparison with other research groups works, both national and international, with possible stages in foreign universities or research centers;

Activity 3.2 analysis of the results obtained by applying the developed methods to the selected domains;

Activity 3.3 writing of the doctorate thesis.

	1st year				**2nd year**				**3rd year**			
Activity	*Q1*	*Q2*	*Q3*	*Q4*	*Q1*	*Q2*	*Q3*	*Q4*	*Q1*	*Q2*	*Q3*	*Q4*
1.1	🟧	🟧	🟧									
1.2		🟧	🟧	🟧								
1.3		🟧	🟧	🟧								
1.4		🟧	🟧	🟧								
2.1					🟦	🟦	🟦					
2.2					🟦	🟦	🟦	🟦				
2.3						🟦	🟦	🟦				
3.1									🟩			
3.2										🟩	🟩	
3.3									🟩	🟩	🟩	🟩

8. Result evaluation

To evaluate the results obtained by performing link prediction, it is possible to use different metrics from literature, such as:

- $Precision = \frac{TP}{TP+FP};$
- $Recall\ (or\ True\ Positive\ Rate) = \frac{TP}{TP+FN};$
- $False\ Positive\ Rate = \frac{FP}{FP+TN};$
- $F - measure = 2 * \frac{Precision*Recall}{Precision+Recall}.$

In the previously formulations, TP stands for True Positives, FP stands for False Positives, TN stands for True Negatives and FN stands for False Negatives.

Another type of evaluation can be performed through graphical plots such as ROC and PR curve. ROC curve stands for Receiver Operating Characteristic curve and its main aim is to show the tradeoff between True Positive Rate and False Positive Rate at different thresholds. The area under the ROC curve (AUROC) depicts the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming "positive" ranks higher than "negative").

Another curve that could be plotted is the Precision-Recall curve. It can be created by plotting Precision against Recall at different threshold. It is possible to calculate the area under the PR curve (AUPRC). AUPRC is useful in those domains where there class imbalance problem occurs.

We cannot use all metrics in all domains because there could be some domains which has lack of some information. It is the case of the biological domain. With this domain, indeed, often it is possible only calculate recall because there is a lack of True Negatives, that is negative associations confirmed.

It can be interesting analyze the evolution of the previously mentioned metrics in conformity with the increase of the available data. To reach this goal, it is possible to use Recall@K, Precision@K and F-measure@k. For example, Recall@k calculates recall value with the first (or the best) k instances of data. It is also possible using these metrics to draw ROC and PR curves and calculate the area under them. For instance, ROC curve could be created by plotting the Recall@K against the FPR@K.

9. Possible reference persons external to the department

I hope to identify some possible reference persons external to the department, working in European universities or European research centers, during summer schools or during time spent abroad.
Nowadays, the possible reference persons external to the department are:

- Michalis Vazirgiannis - LIX, École Polytechnique, Palaiseau, France;
- Aristides Gionis - Department of Computer Science, Aalto University, Aalto, Finland;
- Jilles Vreeken - Cluster of Excellence Multimodal Computing and Interaction, Saarland Informatics Campus, Saarbrücken, Germany.

10. References

[1] S. Fakhraei, J. Fould, M. Shashanka and L. Getoor, "Collective Spammer Detection in Evolving Multi-Relational Social Networks," in *KDD '15*, Sydney, NSW, Australia, 2015.

[2] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz and N. Ramakrishnan, "Capturing Planned Protests from Open Source Indicators," *Ai Magazine,* vol. 37, pp. 63-75, 2016.

[3] A. Ramesh, M. Rodriguez and L. Getoor, "Multi-relational Influence Models for Online Professional Networks," in *International Conference on Web Intelligence*, Leipzig, Germany, 2017.

[4] T. Rekatsinas, S. Ghosh, S. R. Mekaru, E. O. Nsoesie, J. S. Brownstein, L. Getoor and N. Ramakrishnan, "Forecasting rare disease outbreaks from open source indicators," *Statistical Analysis and Data Mining: The ASA Data Science Journal,* vol. 10, no. 2, pp. 136-150, 2017.

[5] S. H. Kumar, J. Pujara, L. Getoor, D. Mares, D. Gupta and E. Riloff, "Unsupervised Models for Predicting Strategic Relations between Organizations," in *International Conference on Advances in Social Networks Analysis and Mining*, San Francisco, CA, USA, 2016.

[6] C. S. Mouli, A. Naik, B. Ribeiro and J. Neville, "Identifying user survival types via clustering of censored social network data," *arXiv,* 2017.

[7] S. H. Bach, M. Broecheler, B. Huang and L. Getoor, "Hinge-Loss Markov Random Fields and Probabilistic Soft Logic.," *Journal of Machine Learning Research (JMLR),* vol. 28, 2015.