My main research interest lies in developing machine learning and large-scale data mining methods for the analysis and modeling of large real-world networks and processes that take place over them. In this context, there is a wealth of research problems and high-impact applications in social networks, information systems, marketing, epidemiology, business intelligence, and national security. For example, identifying influential users in social networks can become a multi billion dollar industry; detecting the spread of rumors and misinformation can improve information reliability and trustworthiness; designing user interfaces that mitigate information overload can increase users' engagement and improve workers' productivity; early detection and control of disease spreads can save lives.

My research consists of several dimensions: (1) developing realistic models of networks and processes that take place over them, assessing their theoretical properties and limitations; (2) developing machine learning algorithms to fit such models and computational methods to influence processes over networks; (3) validating these models and methods in massive real-world datasets of gigabyte and terabyte scale.

In the upcoming years, I would like to focus on building models and algorithmic inference methods to understand, predict, and influence information diffusion over networks. To reach this goal, I carry out a bottom-up approach, which starts by considering particular ideas, pieces of information, decisions, or, more generally, contagions, which spread locally from node to node apparently at random but later, produce global, macroscopic patterns at a network level. To bridge the gap between local dependencies and global patterns, I leverage recent methods from machine learning, probabilistic modeling, event history analysis and graph theory, as well as the nascent field of network science.

Ultimately, I aim to provide computational tools with direct applications in a wide range of domains such as social and information sciences, economics, decision theory, causality or epidemiology.

1. Current Achievements

Observing a diffusion process often reduces to noting when nodes (people, blogs, etc.) reproduce a piece of information, buy a product, get infected by a virus, or, more generally, adopt a contagion. We often observe *where and when* but not *how or why* contagions propagate through a population of individuals. One of the main goals of my dissertation [1] was to shed light on the hidden underlying structure and temporal dynamics of information diffusion, with a particular emphasis on the diffusion of information over blogs and mainstream media sites. To this aim, I developed flexible probabilistic models and inference algorithms that make minimal assumptions about the physical, biological or cognitive mechanisms responsible for diffusion. This is possible since my models are data-driven and rely primarily on the visible temporal traces that the diffusion generates, called *cascades*.

Below, I elaborate on some of the main themes of my thesis research, which I summarize in the following Table.

Section	Ref.	Topics and tools
1.1	[2-4]	Network inference, submodular optimization
1.2	[5, 6]	Network inference & temporal dynamics, convex optimization
1.3	[6, 7]	Dynamic network inference, stochastic convex optimization
1.4	[8]	Signed network inference, convex optimization
1.5	[9, 10]	Influence estimation & maximization, submodular optimization, randomized methods
1.6	[11-13]	Other achievements: survival theory, events & network evolution

Table 1: Publications, topics and tools of my thesis research

1.1 Inferring Diffusion Networks. I was among the first to develop an efficient algorithm, NETINF, to infer the structure of a network from the temporal traces left by information diffusion [2, 3, 4]. The algorithm has provable near-optimal performance based on submodular maximization. NETINF allowed me to study information diffusion among mainstream media and blogs sites. I experimented with more than 170 million blogs and news articles and found that the diffusion network of news for the top 1,000 media sites and blogs tends to have a core-periphery structure with a small set of core media sites that diffuse information to the rest of the Web. These sites tend to have stable circles of influence with more general news media sites acting as connectors between them. This work received a best research paper award honorable mention at ACM KDD 2010 and according to Google Scholar, by October 2013, it is the most cited KDD 2010 paper, accumulating more than 200 citations. KDD (International Conference on Knowledge Discovery in Data Mining) is the flagship conference in data mining.

1.2 Inferring the Temporal Dynamics of Diffusion Networks. Motivated by empirical evidence that indicates that, in real networks, diffusion occurs at different rates across different edges, I developed a new network inference algorithm, NETRATE [5, 6], which allows for heterogeneous rates within a network. The algorithm consists of a convex program, which naturally (without heuristics) imposes sparse solutions and requires no parameter tuning. Remarkably, in spite of the variable transmission rates across edges, the approach ends up being more elegant than NETINF. This work has triggered a significant amount of follow-up work. For example, other researchers have extended NETRATE's model to include textual information, topic modeling and the influence of external events. According to Google Scholar, by September 2013, it is one of the most cited ICML 2011 papers. ICML (International Conference on Machine Learning) is one of the two flagship conferences in machine learning.

1.3 Inferring Dynamic Diffusion Networks. I extended NETRATE's model to support dynamic networks that change over time, depending upon the contagions that propagate through them, and developed INFOPATH [6, 7]. This is important since, for example, a blog can increase its popularity abruptly after one of its posts turns *viral*, creating new edges in the information transmission network and so the content the blog produces in the future will likely spread to larger parts of the network. Although this work has only recently been published (February 2013), it has already attracted a great deal of attention; the companion website received more than 6,000 unique pageviews. By allowing heterogeneous temporal dynamics and dynamic networks, INFOPATH facilitate the discovery of more complex patterns. For example, I found that information pathways for general recurrent topics are more stable across time than for on-going news events. Clusters of news media sites and blogs often emerge and vanish in matter of days for on-going news events. Major social movements and events involving civil population, such as the Libyan's civil war or Syria's uprise, lead to an increased amount of information pathways among blogs as well as in the overall increase in the network centrality of blogs and social media sites.

1.4 Inferring Signed Diffusion Networks. Finally, I also extended NETRATE's model to support signed networks and developed SENTINF [8]. Here, a positive edge between a pair of nodes models consensus or agreement, while negative edges model controversy or disagreement. The algorithm facilitates the study of consensus, controversy and balance in on-line media. For example, I found that both mainstream media and blogs tend to agree more than disagree with their sources and mainstream media sites tend to agree more with their sources than blogs. Depending on the topic, most popular sites can either induce greater consensus or, in contrast, trigger more controversy on what they published. Additionally, the algorithm also helps to automatically identify sites that systematically disagree with the conventional wisdom on different topics. These sites are typically independent media sites, citizen journalism sites, extreme right wing and left wing news media sites, and occasionally blogs with racist, xenophobic, or reactionary views.

1.5 Influencing Diffusion. I applied my probabilistic framework of diffusion to the influence maximization problem. Influence spread maximization tackles the problem of selecting the most influential source node set of a given size in a diffusion network. A diffusion process that starts in such an influential set of nodes is expected to reach the greatest number of nodes in the network. I first developed an influence estimation and maximization algorithm: INFLUMAX [9], which allows for different rates across different edges. In contrast, previous work on influence maximization ignored the underlying temporal dynamics governing diffusion networks — once a transmission occurs, it occurs at the same rate or temporal scale. Experiments on synthetic and real diffusion networks show that INFLUMAX outperforms the previous state of the art.

Then, I developed a highly scalable follow-up randomized algorithm for influence estimation and maximization, CONTINEST [10], which easily scales up to networks of millions of nodes. This work received an outstanding paper award at NIPS 2013. NIPS (Neural Information Processing Systems) is one of the two flagship conferences in machine learning.

1.6 Other achievements. I have also generalized my probabilistic framework of diffusion using survival theory [11, 12] and studied the impact of real-world events on social networks evolution [13].

I believe reproducibility and open knowledge are essential to the advance and dissemination of any field of research. Therefore, I have released companion websites for several of my publications, which typically include software packages, datasets, and additional results.

Ref.	Companion website	Content
[2,3]	http://snap.stanford.edu/netinf	Software package, dataset, additional results
[4]	http://people.tue.mpg.de/manuelgr/nim	Software package
[5, 6]	http://people.tue.mpg.de/manuelgr/netrate	Software package, additional results
[6, 7]	http://snap.stanford.edu/infopath	Software package, dataset, additional results
[9]	http://people.tue.mpg.de/manuelgr/influmax	Software package

Table 2: Companion websites of my publications and their content

2. Previous Achievements

At the beginning of my graduate studies, I applied machine learning and statistical signal processing to the fields of brain machine interfaces (BMI) [14, 15, 16, 17, 18] and computational photography [19]. In the context of BMIs, I developed an integrated stroke rehabilitation framework that combines robot-assisted physical therapy with decoding of movement intent using a brain computer interface (BCI). The rationale is that synchronizing movement intent with robot-assisted therapy may support a cortical reorganization, and thus result in enhanced functional recovery. In the context of computational photography, I developed a machine learning method to denoise astronomical images. These early projects helped me broaden my knowledge of machine learning methods and complex data analysis.

3. Vision for the Future

My long-term research agenda is strongly influenced by complex social, technological and cognitive phenomena that emerge in an increasingly networked digital world. For example, there exist a growing number of natural, social and technological networks over which diffusion takes place. To which extent does information propagate across different networks? Ideas and information, decisions, or, more generally, *contagions* are produced and consumed at increasingly faster rates over increasingly larger social and technological networks. Is there an information processing limits on humans? Networks have become ubiquitous in modern life and both people and technological systems take decisions based on the information that propagates over them. In

this context, the reliability or trustworthiness of information that propagates becomes crucial. Can we automatically identify rumors or misinformation as they start spreading?

In the upcoming years, I will build models and algorithmic inference methods that take into account the above-mentioned phenomena, among others, and help us understand, predict, and influence information diffusion over networks. I will pursue this goal by using a bottom-up approach that first considers how particular contagions spread locally from node to node to later produce global, macroscopic patterns at a network level. My research agenda decomposes into many conceptual problems, which typically span three dimensions: (1) models, (2) algorithms and (3) experimental validation. Each problem is interesting and important in its own right. Here, I briefly discuss five of them.

3.1 Information propagation across multiple networks. I would like to investigate *when* and *how* information propagates among social and technological networks with different characteristics. For example, information has a very short lifetime on microblogging services, while it typically lasts longer on blogs and mass media sites. A piece of information is more likely to be mentioned by a social media user than to get published by a mass media site. However, it is still more likely to get published by a mass media site than by a knowledge network such as Wikipedia. In this context, I will develop models that account for different temporal scales and nodes' information processing capacities across networks. The models will help revealing the roles different networks play in the production, consumption and diffusion of information, as well as assessing how influential each network is.

3.2 Information processing capacity. Since Alvin Toffler popularized the term "Information overload" in his bestselling book Future Shock, it has become a major problem in modern society. Importantly, the advent of social media, online social networking and micro-blogging services has accelerated dramatically the amount of information a person is exposed to, increasing the information overload. In this line of research, I would like to quantify the amount of information overload a person is actually suffering and develop methods to mitigate this problem. Here, I believe it is key to investigate the existence of an information processing limit on humans. My long-term goal will be to develop adaptive algorithms that automatically optimize the flow of information a person is exposed to, enhancing her user experience, increasing her engagement, and *turning down the noise*.

3.3 Information reliability. Another important aspect in the context of propagation over networks is the reliability or trustworthiness of the information that spreads and mutates. For example, are rumors systematically spreading across the same untrustworthy sites? Can we automatically identify rumors or misinformation as they start spreading? How do nodes assess the reliability of a piece of information? I plan to develop models and methods to automatically quantify the reliability of a piece of information that propagates over a network and the trustworthiness of the nodes in the network. This will have many direct applications on fields as diverse as emergency response, news recommendations systems, question-and-answer websites or business intelligence, among many others.

3.4 Network evolution. I will investigate to which extent the evolution of natural, social and technological networks is driven by the information that propagates over them. For example, a Twitter user often starts following another user when they find one of her tweets is interesting. A blog can become popular once one of its posts turns viral, and subsequently will attract more attention (and links) from other sites on the Web. By understanding this aspect, we may be able to design more effective link prediction algorithms and recommendation systems. In the long-term, I would like to develop computational techniques to influence the growth of a network in order to facilitate the communication and exchange of information between nodes.

3.5 Influencing information propagation. Beyond passive observational studies, which allow us to understand the past and make predictions about the future, I will also carry out active interventional studies, in which we influence current and future information diffusion. To this aim, I will build on previous research done in causality and randomized experiments and perform interventions in both the networks and the diffusion mechanisms. The implications of this line of

research will span many important applications. It will help to prevent the spread of misinformation and rumors, mitigate information overload and its effects, design more effective vaccination and quarantine policies, develop viral marketing campaigns, or design *better* user interfaces, among many others.

The above-mentioned dimensions will require an evolving set of unique computational methods leveraging large-scale machine learning, probabilistic modeling and graph theory, as well as the nascent field of network science. Moreover, it will provide computational tools with direct applications in a wide range of domains such as social and information sciences, economics, decision theory, causality, and epidemiology. I am excited about the influence my graduate research has already had on the fields of machine learning, data mining and network science and looking forward to continue making an impact on both theoretical foundations and real-world applications.

References

[1] **M. Gomez Rodriguez**. Structure and Dynamics of Diffusion Networks. *Ph.D. Thesis, Department of Electrical Engineering, Stanford University*, June 2013.

[2] **M. Gomez Rodriguez**, J. Leskovec and A. Krause. Inferring Networks of Diffusion and Influence. *Proc. of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD**)*, 2010 (**Best Research Paper Award Honorable Mention**).

[3] **M. Gomez Rodriguez**, J. Leskovec and A. Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (**TKDD**)*, Volume 5, Number 4, 2012.

[4] **M. Gomez Rodriguez** and B. Schölkopf. Submodular Inference of Diffusion Networks from Multiple Trees. *Proc. of the 29th International Conference on Machine Learning (**ICML**)*, 2012.

[5] **M. Gomez Rodriguez**, D. Balduzzi and B. Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks. *Proc. of the 28th International Conference on Machine Learning (**ICML**)*, 2011.

[6] **M. Gomez Rodriguez**, J. Leskovec, D. Balduzzi and B. Schölkopf. Uncovering the Structure and Temporal Dynamics of Information Propagation. *Network Science (**NWS**)*, accepted.

[7] **M. Gomez Rodriguez**, J. Leskovec and B. Schölkopf. Structure and Dynamics of Information Pathways in Online Media. *Proc. of the 6th International Conference on Web Search and Data Mining (**WSDM**)*, 2013.

[8] **M. Gomez Rodriguez**, J. Leskovec and B. Schölkopf. Inferring Signed Networks of Diffusion. In preparation.

[9] **M. Gomez Rodriguez** and B. Schölkopf. Influence Maximization in Continuous Time Diffusion Networks. *Proc. of the 29th International Conference on Machine Learning (**ICML**)*, 2012.

[10] N. Du, L. Song, **M. Gomez Rodriguez** and H. Zha. Scalable Influence Estimation in Continuous Time Diffusion Networks. *Advances in Neural Information Processing Systems (**NIPS**),* 2013 (**Outstanding Paper Award**).

[11] **M. Gomez Rodriguez**, J. Leskovec and B. Schölkopf. Modeling Information Propagation with Survival Theory. *Proc. of the 30th International Conference on Machine Learning (**ICML**)*, 2013.

[12] **M. Gomez Rodriguez** and B. Schölkopf. Modeling Information Propagation with Survival Theory. *Advances in Neural Information Processing Systems: Workshop in Algorithmic and Statistical Approaches for Large Social Networks (**NIPS**)*, 2012.

[13] **M. Gomez Rodriguez** and M. Rogati. Bridging Offline and Online Social Graph Dynamics. *Proc. of the 21st ACM International Conference on Information and Knowledge Management (**CIKM**)*, 2012.

[14] **M. Gomez Rodriguez**, J. Peters, J. Hill, B. Schölkopf, A. Gharabaghi and M. Grosse-Wentrup. Closing the sensorimotor loop: haptic feedback facilitates decoding of motor imagery. *Journal of Neural Engineering (**JNE**)*, Volume 8, Number 3, 2011.

[15] **M. Gomez Rodriguez**, M. Grosse-Wentrup, J. Hill, A. Gharabaghi, B. Schölkopf and J. Peters. Towards Brain-Robot Interfaces in Stroke Rehabilitation. *Proc. of the 12th International Conference on Rehabilitation Robotics (**ICORR**)*, 2011.

[16] **M. Gomez Rodriguez**, J. Peters, J. Hill, B. Schölkopf, A. Gharabaghi and M. Grosse-Wentrup. Closing the Sensorimotor Loop: Haptic Feedback Facilitates Decoding of Arm Movement Imagery. *SMC Workshop in Shared-Control for BMI (**SMC**)*, 2010.

[17] **M. Gomez Rodriguez**, M. Grosse-Wentrup, J. Peters, G. Naros, J. Hill, B. Schölkopf and A. Gharabaghi. Epidural ECoG Online Decoding of Arm Movement Intention in Hemiparesis. *ICPR Workshop on Brain Decoding (**ICPR**)*, 2010.

[18] **M. Gomez Rodriguez**, J. Peters, J. Hill, A. Gharabaghi, B. Schölkopf and M. Grosse-Wentrup. BCI and robotics framework for stroke rehabilitation. *4th International BCI Meeting*, 2010.

[19] **M. Gomez Rodriguez**, J. Kober and B. Schölkopf. Denoising photographs using dark frames optimized by quadratic programming. *Proc. of the 1st IEEE International Conference in Computational Photography (**ICCP**)*, 2009.