DR. GONZALO MATEOS

Carnegie Mellon University
Computer Science Department
5000 Forbes Avenue
Pittsburgh, PA 15213

Tel: (612) 859-5059 (Mobile)
Tel: (412)-268-2000 (Office)
Email: mateosg@cs.cmu.edu
http://www.cs.cmu.edu/~mateosg/

Research statement

The overarching theme of my research is on algorithms, analysis, and application of statistical signal processing tools to dynamic network health monitoring, social, power grid, and *Big Network Data* analytics. Experience, background, and current research include robust, distributed, and sparsity-aware learning from high-dimensional network data; which finds exciting applications in understanding the dynamics of emergent social-computational systems, and protecting critical infrastructure such as the Internet's backbone. My long-term research objective is to harness the power of (graph) data to bring significant science and engineering advances of networked systems, along with consequent improvements in quality of life.

Robust statistical learning from 'Big Network Data'

Context, motivation, and current challenges. The information explosion propelled by the advent of online social media, the Internet, and the global-scale communications has rendered *statistical learning from data* increasingly important. At any given time around the globe, large volumes of data are generated by today's ubiquitous and increasingly interconnected communication and mobile devices such as cell-phones, surveillance cameras, e-commerce platforms, and social-networking sites.

The term 'Big Data' is coined to precisely describe this information deluge. Quoting a recent article published in The Economist 'The effect (of Big Data) is being felt everywhere, from business to science, and from government to the arts'. Undeniably, tremendous economic growth and improvements in quality of life can be effected by harnessing the benefits of analyzing such large volumes of data. Making sense of modern massive-scale datasets will facilitate learning the behavioral dynamics of emergent cyber-physical and social-computational systems, as well as protect critical infrastructure including the smart grid and Internet's backbone network. But great promises come with formidable research challenges; as Google's chief economist explains in the same article 'Data are widely available, what is scarce is the ability to extract wisdom from them.' While significant progress has been made in the last decade towards achieving the ultimate goal of 'making sense of it all,' the consensus is that we are still quite not there.

A major hurdle is that massive datasets from e.g., the Internet are often noisy, incomplete, vulnerable to cyber-attacks, and prone to *outliers*, meaning data not adhering to postulated models. Resilience to outliers is of paramount importance in fundamental statistical learning tasks such as prediction, model selection, and dimensionality reduction. In particular, the workhorse least-squares (LS) method is known to be very sensitive to outliers. My Ph.D. research deals with *robust* learning algorithms that are resilient to outliers, and contributes to the theoretical and algorithmic foundations of Big Data analytics.

Doctoral thesis research at-a-glance and contributions. Modern data sets typically involve a large number of attributes. This fact motivates predictive models offering a *sparse*, meaning parsimonious, representation in terms of few attributes to facilitate interpretability. Recognizing that outliers are sporadic, results in my

dissertation establish a neat link between sparsity and robustness against outliers, even when the chosen models are not sparse. Leveraging sparsity for model selection has made headways across science and engineering in recent years, but its ties to robustness were so far largely unexplored. A main contribution is that controlling sparsity of model *residuals* leads to statistical learning algorithms that are *scalable* (hence computationally affordable), and *universally robust* to outliers. Focus is henceforth placed on robust principal component analysis (R-PCA), a novel computational method I devised to extract the most informative low-dimensional structure from (grossly corrupted) high-dimensional data. Scalability is ensured by developing low-complexity streaming and distributed algorithms that build on recent advances in convex optimization, which also offer (analytically) quantifiable performance. Universality of the developed framework amounts to diverse generalizations not confined to specific: i) probability distributions of the nominal and outlier data; ii) nominal (e.g., linear) predictive models; and iii) criteria (e.g., LS) adopted to fit the chosen model.

Major results in my Ph.D. thesis impact everyday life in human-computer interactions, including online purchases aided by e.g., Amazon's automated recommendations. These suggestions could be biased however, even if only a few negative product ratings introduced by 'social liars', or malicious actors with commercially-competing interests, go undetected. The low-dimensional data model of (possibly corrupted) ratings advocated by R-PCA offers a natural fit, since it is now well accepted that consumer preferences are dictated by just a few unknown latent factors. In a related context, remarkable results were obtained in collaboration with faculty from the Dept. of Psychology, when using R-PCA to reveal aberrant or otherwise invalid responses in web-based, self-administered personality assessment surveys. This addresses a timely pressing issue in psychometrics, since 'workhorse methods' used for traditional pencil-and-paper questionnaires are now deemed overly stringent and computationally intensive for online data screening.

My research has also focused on robust learning from *network data*. In this context, traffic spikes in the Internet, power consumption glitches across the power grid, or, fraudsters and agents exhibiting abnormal interaction patterns with peers in social networks can be all reinterpreted as outliers. Much of the challenge in analyzing network data stems from the fact that they involve quantities of a relational nature. As such, measurements are typically both high-dimensional and dependent. Distinct from prior art, the developed algorithms can process network information streams in real time, and are flexible to cope with decentralized data sets that cannot be communicated and processed centrally due to security, privacy, or memory constraints. Leveraging these online and decentralized features, yields an innovative approach to unveiling Internet traffic anomalies using real-time link utilization measurements. The novel ideas here are to judiciously exploit the low-dimensionality of end-to-end network flows and the sparse nature of anomalies, to construct a map of anomalies across space and time. Such a powerful monitoring tool is a key enabler for network health management, risk analysis, security assurance, and proactive failure prevention. The proposed algorithms markedly outperform existing approaches, that neither account for the underlying graph (they focus on isolated link measurements), nor they capitalize on the sparsity of anomalies.

Future research agenda and funding plans

In my future research, I plan to continue building upon the theoretical and algorithmic foundations developed during my doctoral study, as well as to expand towards new directions at the intersection of *Big Data* research and *Network Science*. I am convinced that sizable impact can be made by devising novel encompassing models applicable to a wide range of network analytics problems, and offering architectures and algorithms to handle the practical challenges, while revealing fundamental limits and insights regarding the various statistical trade-offs involved.

A particular area of interest is to explore Big Data *tensors* (also known as multi-way arrays) for representation and synthesis of dynamic network data, e.g., a large who-calls-whom social graph observed over a given time horizon. Unveiling low-rank structures in tensors is well-motivated to spot e.g., bipartite cores and large stars due to telemarketers. Albeit natural this can be a challenging proposition, since even computing a tensor's rank is itself an NP-hard problem. Adopting a convex surrogate for promoting low rank such as the nuclear norm in the matrix case is not obvious either, since singular values counterparts are not related with the notion of tensor rank. Inspired by R-PCA where a redefinition of the (matrix) nuclear norm is obtained in terms of the Frobenius norms of the low-rank matrix factors, I will investigate novel techniques for low-rank tensor decomposition and imputation under various computational and data-related challenges. Is the decomposable structure of such models amenable to MapReduce/Hadoop implementations scalable to million-node graphs? Streaming algorithms are also desirable since Big Data problems often come with time constraints, where a high-quality answer that is obtained slowly can be less useful than a medium-quality answer that is obtained quickly. Intriguing theoretical questions can be posed as well. Aiming at a general approach to anomaly detection from graph data, are the envisioned low-rank plus sparse tensor decomposition models identifiable? Latent factors of the low-rank components should be also sparse, since only few members of each mode belong to a community. Is there enough structure to discern interesting cliques from the random patterns of sparse anomalies? Is there a principled way to explicitly model the temporal evolution of communities, and thus enhance the discriminative power of the algorithms.

Funding experience and resources. External funding is important to conduct research and support graduate students. During my Ph.D. I have contributed in drafting various proposals and whitepapers, several of which involved a cross-disciplinary team across the computational and social sciences, as well as joint efforts with industry partners. I also have extensive technical collaboration experience with faculty and colleagues from computer science, psychometrics, and marketing, which I value as an important first step towards overcoming a potential 'language barrier'.

In terms of funding programs, the importance of Big Data research nowadays is apparent. On 2012, the White House Office of Science and Technology Policy announced the Big Data Research and Development Initiative. The launch includes over $200 Million in new commitments through the National Science Foundation (NSF), National Institutes of Health, Department of Defense (DoD) at large, Department of Energy, and the US Geological Survey. A first noteworthy example is the NSF/CISE 12-499 program, Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA). The main goal of this program is '...to extract and use knowledge from collections of large data sets in order to accelerate progress in science and engineering research.' In addition, DoD is placing a 'Big Bet on Big Data' with more than 20 open solicitations. I concur with the basic premise of a recent NSF-sponsored Big Data workshop (http://www.dtc.umn.edu/bigdata/), namely that the signal processing and systems engineering communities can be important contributors to Big Data research and development, complementing computer and information science-based efforts in this direction.

As a junior faculty I will apply for the NSF CAREER Award in July 2015, and will actively seek external funding opportunities through suitable channels. The research described in this statement fits the interest of at least four programs of the NSF: the aforementioned BIGDATA solicitation, Computing and Communication Foundations (CCF) – closing on September 19, 2014 – Communications, Circuits, and Sensing-Systems (CCSS) – with deadline on January 29, 2015, and Cyber-Physical Systems (CPS) in which '...physical processes are tightly intertwined with *networked* computing'. The Defense Advanced Research Projects Agency (DARPA) released a document describing 23 Mathematical challenges. Challenge number two is 'The Dynamics of Networks', and shares my interest to 'accurately model and predict behavior in

large-scale distributed networks that evolve over time occurring in communication, biology, and the social sciences'. The Air Force Office of Scientific Research (AFOSR) is currently funding the research described here; my M.Sc. Thesis work on distributed information processing using wireless sensor networks was supported by Army Research Lab (ARL).