

# Research statement

Tomas Singliar

tomas@cs.pitt.edu

Computer Science Department, University of Pittsburgh

July 22, 2008

## Abstract

This document is filled with questions that I aspire to answer in my future research. The main areas of interest are stochastic network systems, event detection and robust learning from noisy, real-world data. All are bound together by the powerful methodology of statistical machine learning.

Many real-world problems do not have a clear, feasible algorithmic solution. For most of these problems, machine learning inspires an approximate solution that is, surprisingly often, “good enough”. The machine learning attack is to generalize from collections of instances of similar problems. A class of such problems that is especially challenging involves complex connected systems. These networks underpin our lifestyle: transportation, distribution and communication. A continuum of fascinating problems involving modeling, prediction and decision making naturally emerges when one studies their behavior. Can we predict road traffic patterns? How much should an insurance company charge for insuring against damages caused by an electrical service outage? How is a disease or a new computer virus likely to spread and how do we stop it fast? These problems share a complex underlying link structure and attempts to crack them with statistical tools have generated a wealth of theoretical questions in Machine Learning for me, as well an appreciation for the challenges of working with real-world data.

**Stochastic networks.** Consider an everyday task for many of us—choosing a driving route to work. Devices now exist that can help us with the navigation task, even providing up-to-minute traffic information. But ideally, we would like more than that: we would like to base our planning on what traffic situation we will encounter in the future, after we have traveled for many minutes. I proposed a traffic flow model that gives a distribution over the *future* state of traffic. It does so by marrying models of traffic flow to probabilistic inference techniques, especially particle filtering.

I have also shown that having such prediction can indeed improve routes suggested by the planner under the expected travel time criterion, especially in volatile and congested situations [9]. However, the expected travel time may not always be the ideal criterion. For instance, minimizing the probability of being late is a reasonable objective. *Optimization with such objectives is underexplored and an alluring direction of future work.*

The landscape of Intelligent Transportation Systems is being changed by ubiquitous mobile sensing. If many vehicles are equipped with wireless-networked GPS devices, there is no longer

need to install, at great expense, a few sensor stations. The “commodity flow” *is* the sensor! New vistas open for traffic flow modeling with this development. No longer must a model discretize a roadway into chunks defined by sensor placement. *New modeling paradigms will be necessary to accommodate these new formats of data.*

An early-adopted and important component of Intelligent Transportation systems is automated incident detection (AID). I proposed a Machine Learning approach using SVMs to AID [7] that has the advantage of being self-adapting to data.

In the long run, performance of networked systems could be optimized globally from a central location collecting much of the information. Even contemplating such idea brings hosts of game theoretic, computational, economic, privacy and ethical questions: *Can we agree on the definition of social welfare? How do we optimize it computationally when we find the welfare function? Can this data be collected in a privacy-preserving way?*

Understanding of the complex stochastic behavior of networks enables more accurate solution of many common operations research problems. In [2], we considered a problem of choosing a supplier of a commodity, where the choice is affected not only by price, but also by the reliability of the network through which the commodity is to be delivered. Importantly, this paper was among few that proposed a tool to cope with correlated, cascading link failures. *However, how do we know we got the failure model right?* It is difficult to know, since there is (luckily!) so little data. This is a feature that study of stochastic networks shares with anomaly detection: very little positive instances of data, prompting the question: *How does one induce and validate models on extremely skewed datasets?*

**Event and anomaly detection.** The collaboration with the Distributed Detection and Inference group at Intel Research has been very fruitful and spurred in me an interest in automated detection of rare but significant events, such as disease outbreaks in human populations, malware intrusions in computer networks [3] or vehicular incidents [7].

The fields of anomaly detection and machine learning interlock and enrich each other. In [7], our most challenging issue was to develop a learning technique to remove systematic biases in the incident data using dynamic Bayesian networks (DBN). A “trick” to speed up DBN inference was also independently discovered by my collaborator on malware intrusion detection [3, 1]. This led us to examine the properties of the transformation and we found a linear algorithm for general inference in a certain class of dynamic Bayesian networks [4] that shows up often in detection and monitoring tasks.

Detection of epidemic outbreaks is a major concern of public safety. Despite much recent work in automated detection from data streams, currently the detection of disease outbreaks often relies on a human eye. Maybe this is the way to go—people are notoriously good at spotting anomalies. *Are there better visualizations of outbreak data that take advantage of how humans perceive information? Can we give better up-to-date live maps to public health officials?* Clearly, collaboration with cognitive scientist would be needed to answer such questions.

The issue of evaluating (anomaly) detectors is in my opinion absolutely critical to further progress in this domain. The machine learning community understands quite well what it means to say that a classifier performs well. I don’t think we can say the same about detectors. Sensitivity and specificity may not be enough: For instance, how often does the detector evaluate its input signal affects these metrics, time-to-detection and other performance met-

rics. Are we safe to assume that closely spaced detections are uncorrelated? *We need a better understanding of how to evaluate event detection algorithms.*

**Robust learning with structured, noisy data.** In my work on traffic prediction and planning, I encountered another major challenge that inspires my research agenda, besides the inherent complexity of the transportation network: the complexity of dealing with real world data. Machine learning theory usually thinks about the data as a datapoint-by-attribute matrix. Unfortunately, data does not look like that in the real world. Real data comes with semantics: road traffic sensors are pinned to a geographic location and produce data for each of the varying number of lanes. Real data comes on varying time scales. Real data does not go missing uniformly at random and requires cleaning. To enable inference of missing data, I characterized the bird’s eye vies of traffic flow patterns [8] by a generative probabilistic model. While designed for traffic data, it is a general-purpose model applicable to any dataset with continuous values. As exemplified by this work and the paper on incident detection [7], *even the simple need to regularize data gives rise to general sophisticated methods* with wider consequences for the theory and practice of machine learning.

**Data mining.** Social networks are a fashionable example of a virtual interlinked system. The amount of available social network data is currently outpacing our knowledge of its structure. The research area of automated link and community analysis is defined by trying to understand them and transform the information into knowledge. The powerful formalism of graphical models is sure to play a role. For instance, we have developed an approximate learning algorithm for a graphical model that identification of citation communities in CiteSeer data [6, 5]. Th model structure embodies the assumption that citation each document can be represented by a set of active “topic factors”. *My intuition is that the structure of a graphical model that succesfully learns to perform a task in a linked environment tells us a great deal about the link structure itself.*

## References

- [1] Denver Dash, Branislav Kveton, John Mark Agosta, Eve Schooler, Jaideep Chandrashekar, Abraham Bachrach, and Alex Newman. When gossip is good: Distributed probabilistic inference for detection of slow network intrusions. In *Proceedings of AAAI 2006*, 2006.
- [2] Miloš Hauskrecht and Tomáš Šingliar. Monte-carlo optimization for resource allocation problems in stochastic network systems. In *Proceeding of Conference on Uncertainty in Artificial Intelligence, UAI2003*, 2003.
- [3] Tomáš Šingliar and Denver H. Dash. Cod: Online temporal clustering for outbreak detection. In *Proceedings of 22<sup>n</sup>d Conference on AI, AAAI’07*, 2007.
- [4] Tomáš Šingliar and Denver H. Dash. Efficient inference in persistent dynamic Bayesian networks. In *Proceedings of the 24<sup>th</sup> Conference on Uncertainty in Artificial Intelligence, UAI-2008*, page tbd, 2008.

- [5] Tomáš Šingliar and Miloš Hauskrecht. Variational learning for noisy-or component analysis. In *Proceedings of SIAM International Conference on Data Mining, SDM2005*, pages 370–379, 2005.
- [6] Tomáš Šingliar and Miloš Hauskrecht. Noisy-or component analysis and its application to link analysis. *Journal of Machine learning Research*, pages 2189–2213, Oct 2006.
- [7] Tomáš Šingliar and Miloš Hauskrecht. Learning to detect adverse traffic events from noisily labeled data. In *Proceedings of Principles and Practice of Knowledge Discovery in Databases PKDD 2007*, number 4702 in LNCS, pages 236–247. Springer, 2007.
- [8] Tomáš Šingliar and Miloš Hauskrecht. Modeling highway traffic volumes. In *Proceedings of European Conference on Machine Learning ECML 2007*, number 4701 in LNCS, pages 732–739. Springer, 2007.
- [9] Tomáš Šingliar and Miloš Hauskrecht. Approximation strategies for routing in dynamic stochastic networks. In *Proceedings of the 10<sup>th</sup> International Symposium on Artificial Intelligence and Mathematics—ISAIM 08*, page tbd, January 2008.