Research Statement — Alex Beutel

Can we model *how* fraudsters work to distinguish them from normal users? Can we predict not just which movie a person will like, but also *why?* How can we find *when* a student will become confused or *where* patients in a hospital system are getting infected? How can we effectively **model large attributed graphs** of complex interactions?

As "Big Data" has become pervasive, organizations in every industry are storing as many actions and interactions as they can. For both academia and industry these massive databases have significant untapped potential, with increasing investment in trying to discover useful patterns. These databases can all be viewed as large hypergraphs — graphs among entities of different types with many attributes contextualizing their interactions. Through modeling these interactions, computer science can provide insights beyond what we can see and understand individually and propel many fields forward. With more information being stored than ever before, we now have the opportunity to move beyond only predicting whether two entities have interacted but also **understanding the context of those interactions** based on the many meaningful attributes available. This rapid expansion in the types of interactions and contextual information being stored presents a new frontier for graph modeling, enabling new applications and presenting new challenges in data mining and scalable machine learning.

To give some examples: online users interact not just with each other in social networks, but also with the world around them — supporting politicians, watching movies, buying clothing, searching for restaurants and even finding doctors. These interactions often include insightful contextual information as attributes, such as the time or location of the interaction and ratings or reviews about the interaction. There are similar hypergraphs in healthcare interactions among patients, doctors, diseases and treatments, as well as in educational interactions among students, teachers, subjects and educational resources.

In each of these fields, researchers have been able to extract useful knowledge just from the graph structure, e.g., predicting what movies a person will like and finding communities of people with similar interests. However, to provide more insightful understanding, we need to model not just the interactions but also the context of the interactions, finding the meaningful attributes within a graph and imbuing our models of those attributes with our intuition. Making use of all of these data effectively presents many exciting research challenges, ranging from designing models that capture the relevant patterns in large, attributed hypergraphs to systems for scaling learning of these complex models. Focusing on where classic techniques do not meet real-world challenges, we can develop holistic solutions that **maximize real-world impact**. To do this, my research **bridges insights from applications, modeling and scalable machine learning systems**.

Thesis Research

My thesis work has focused on modeling online user behavior, in particular developing new mathematical and computational techniques where classic techniques fail:

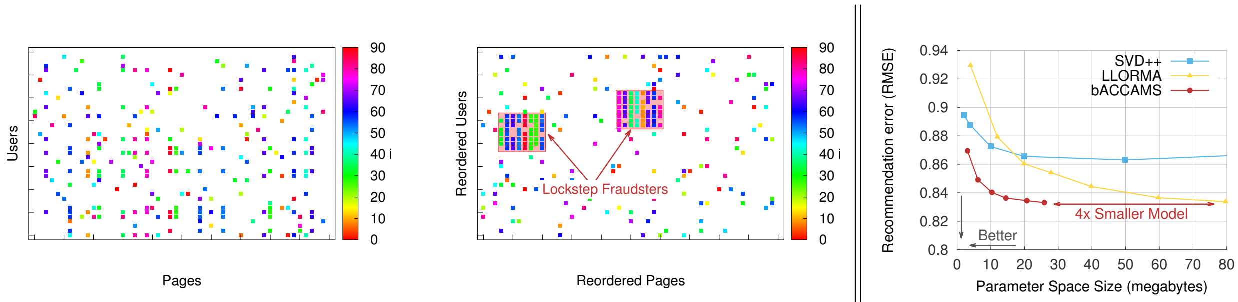
1. **Modeling Abnormal Behavior:** How can we detect fraudulent user behavior? How can we consider the economic incentives of fraud during anomaly detection?
2. **Modeling Normal Behavior:** How can we understand a user's preferences and predict their future actions? How can we explain our predictions?
3. **Scalable Machine Learning:** How can we efficiently distribute learning of these complex models over billions of nodes and actions?

I discuss examples of my work in each of these areas below.

Modeling Abnormal Behavior

Fraud on services like Amazon, Facebook and Twitter deceives honest users, disrupts our models of normal behavior and erodes the value of these services. Therefore, it is crucial that we **detect and remove fraud**. From a graph perspective, fraud is a set of paid edges; for example, fake Page Likes are paid edges in the "who-Likes-what" graph of Facebook, and fake reviews are paid edges in the "who-rates-what" graph of Amazon. Previously, we found fraud of this type by searching for unusually dense subgraphs [8, 12, 10, 9]. However, since honest communities also create dense subgraphs, they are difficult to distinguish from fraudulent behavior using only graph structure. Therefore, I have focused on using contextual information that leaves **distinguishing patterns that are *costly* for fraudsters** to avoid [6, 14, 7, 15].

In [6] we developed an algorithm, CopyCatch, that detects temporally-coherent dense subgraphs or *lockstep behavior* — groups of users who perform the same action at approximately the same time. A visual demonstration of CopyCatch can be seen in Figure 1(a-b). Because fraudsters are obligated to deliver their product (fraud) in a



(a) Before CopyCatch (b) After CopyCatch (c) ACCAMS

Figure 1: Examples of modeling interaction networks for fraud detection (a-b) and recommendations (c).

short timespan, CopyCatch is difficult for adversaries to avoid. Furthermore, CopyCatch has high precision because normal users' actions are often uncorrelated with time. The CopyCatch implementation can **distribute learning** over a thousand machines, searches over **billions of Page Likes** and was **relied on by Facebook** for years to detect Page Like fraud.

In [7] we developed new techniques for incorporating additional costly features, such as IP addresses, for finding spam and fraud. Our flexible search tool can take many different components of user behavior and find which components are most indicting for a particular group of users. Together, [6, 14, 7, 15] have found previously undetected hijacked users adding fake Page Likes on Facebook, fake followers on Twitter, hashtag boosting and purchased retweets on Tencent Weibo, and fake reviews on Flipkart.

Impact:

- CatchSync [8] was a **Best Paper Award finalist** at *KDD* 2014, improving over state-of-the-art by up to 36%.
- CopyCatch [6] was **used at Facebook** for years and inspired multiple follow up projects there.
- CopyCatch [6] was published in *WWW* 2013 and **patented** by Facebook (granted July 7, 2015).
- CopyCatch [6] received an **Editor Highlight** in the ACM Computing Review.

Modeling Normal Behavior

Online, we can interact with each other and the world around us at an unprecedented scale. Through modeling user behavior, we can understand these interactions and help sort through the deluge of options available online. With these goals, I have developed theoretical models to **understand viral content** [13, 4] and improved data-driven models of user preferences for **recommendation systems** [3, 1, 16]. In both sets of research, I have designed novel models to better **match real-world constraints**.

Because recommendation systems have been based on general matrix models, they make the implicit assumption that ratings are Gaussian — that is, there is some "correct" rating and each user gives that rating with some slight personal perturbation. However in [3] we find that, in practice, many items are given highly skewed or bimodal reviews, where nearly all reviews have 1-star or 5-stars. Therefore, we designed a more flexible model and novel sampler that learns both Gaussian and bimodal rating distributions. Our model, CoBaFi, more accurately captures rating preferences and allows us to find polarizing items or risk-averse users.

While recommendation accuracy is important, users are more likely to follow these recommendations if they are accompanied by an explanation. Thus, having simple, **interpretable models** is highly valuable. Previous factorization-based approaches create large models that are difficult to explain. In [1] we take a radically different approach to recommendation by learning a small set of discrete latent attributes about users and items, and the predicted ratings that accompany these attributes. That is, we learn an additive co-clustering model, ACCAMS, of the user and item ratings matrix. As seen in Figure 1(c), ACCAMS achieves state-of-the-art prediction accuracy on Netflix with a model $1/4$ of the size. Recently in [16], we offer an additive Poisson co-clustering model, enabling us to jointly model ratings and text reviews and to **explain our recommendations** with words.

Across all of this research, the models are designed to better understand attributes (ratings and reviews). As a result, these models are both more accurate and help explain *why* a user will like a particular item.

Impact:

- ACCAMS [1] achieves **state-of-the-art accuracy at $1/4$ of the model complexity**.
- Since it was published in *WWW* earlier this year, ACCAMS [1] **open-source code** has been downloaded over 150 times from 14 universities and over 20 countries.

Scalable Machine Learning

To scale complex behavior models to the high volume and variety of data present online, my research efficiently **distributes learning** by exploiting the structure of our models and the stochastic nature of learning [2, 11, 5]. While some methods above, such as CopyCatch [6], include custom implementations that scale to large graphs, many other methods are based on more common latent factor structure. Therefore, I have focused on effectively learning **massive latent factor models**.

In [2] we examine how to learn latent factor models for complex datasets where entities have relations of multiple types, represented as tensors and joined tensors. In order to scale, we need to handle not just "big data" but also learn a model for many entities with complex interactions. We develop FlexiFaCT to scale learning of huge models, with **billions of parameters**, from big datasets. FlexiFaCT scales effectively by understanding the intrinsic structure and independence assumptions of these models and thus partitioning both data and model across many machines.

Distributed machine learning often relies on unreliable clusters, with concurrent programs causing slow machines or even high-demand machines being preempted. In [11] we offered a novel solution for fast machines to not waste time when waiting for slower machines, called *stragglers*. By exploiting the stochastic nature of machine learning, our system is up to **25-times faster** than competing methods. Later, in [5] we further developed our factorization model, in this case for Bayseian models, and offered a technique for **elastic learning**, where the work can be dynamically scaled based on the number of machines that are available.

Each of these solutions exploits the particular structure of relational models and the unique properties of machine learning algorithms to address often frustrating realities of learning these models in the real world. By focusing on the particular properties of learning this broad class of models we develop significantly improved solutions.

Impact:

- FlexiFaCT [2] is the **most cited paper** of *SDM* 2014[1].
- Taught in Carnegie Mellon's graduate course "Machine Learning with Large Datasets" both in 2014 and 2015.

Future Research

Through modeling large attributed graphs we can model and predict not just "who" and "what," but also "why," "how," "when," and "where." In my thesis research I have improved the state-of-the-art in user behavior modeling: modeling how fraud is created, offering explainable recommendations, and creating systems for scaling the learning of complex behavior models. In the coming years, I plan to take a multipronged approach to making hypergraphs broadly useful: (1) solve modeling challenges in **new applications**, focused on where our prior research can have the greatest impact and where new tools need to be developed, (2) develop **new modeling techniques** that address the practical limitations of classic modeling techniques but are general enough to be used across a wide variety of applications, (3) develop **systems for scalable learning** of models of large, complex networks. Through bridging these three challenges, we can develop new insights and holistic solutions to maximize real-world impact. I discuss the research opportunities and challenges for each of these areas below.

Application Outreach: My first goal is to expand the use of graph modeling to new applications. As described above, all relational databases can be considered large hypergraphs. As a result, there are fascinating opportunities in healthcare, education, security, distributed systems, financial systems, and many other fields. For example, in **education** can we predict what resources make a student enter or leave a field? What factors cause a student to misunderstand a concept, and what resources resolve that confusion? In **healthcare**, what combinations of unrelated diseases or treatments develop together? Can we predict when a treatment plan will fail, based on the circumstances of the patient? Through modeling graphs, we can change the way these fields develop, enabling new types of questions to be asked and giving actionable answers. I plan to closely **collaborate** with researchers and practitioners in these fields to understand the unique constraints of their problems and develop custom, useful solutions.

Modeling and Algorithms: The second challenge is in developing **general techniques for understanding attributed hypergraphs**. For example, how do we find which relation types and attributes are important? At a higher level, parts of the graph, edges or attributes, may have relevant meaning that we want to include. How can we include the insights of economics or biology in our models? To be concrete and to generalize on my thesis research: how can we **include arbitrary incentive functions and economic models**, such as the cost and profit opportunities for fraudsters, in our models of graphs? In other fields, can we find behavior that overfits

[1] Based on `https://scholar.google.com/scholar?cites=7979229435915048938`, as of October 29, 2015.

to economic incentives rather than predictive models, such as risk-averse doctors over-medicating or over-testing? Designing algorithms to understand attributed graphs and incorporate our intuition on these attributes will lead to more accurate and powerful models.

Scalable Machine Learning: The strength and opportunity in modeling graphs comes from the ability to jointly understand **many complex interactions**. Therefore, by improving the scalability of our learning algorithms and models, we can directly increase the impact of these models. In scaling graph models, there are multiple ongoing challenges: How can we learn more complex models, of more entities and more attributes, with more data? How can we learn these models quickly? When should we trade off accuracy for speed? Building **scalable, flexible, and expressive systems** for modeling large complex hypergraphs will improve the development of **new models** and the usefulness of those models in practice.

Collaboration, Data and Funding: Throughout my Ph.D. work, I have developed and fostered successful collaborations across many institutions, including **Facebook, Microsoft, Google, Yahoo, and Flipkart**. These fruitful collaborations have provided **rich datasets** and valuable insight from practitioners, resulting in high quality research [6, 5, 15]. This approach to research has been highly successful in attracting both government and industry funding. I have been funded by a **Facebook Fellowship** and an **NSF Graduate Research Fellowship**. Additionally, I have worked closely with my advisor to obtain additional funding through Yahoo's Faculty Research and Engagement Program (FREP) and an NSF collaborative grant (award number IIS-1408924).

In summary, my focus has been and will be to design scalable techniques for modeling large, attributed graphs. My approach to research is to maximize impact through focusing on where classic techniques do not meet modern, real-world challenges. In this effort, my research bridges insights from applications, modeling and scalable machine learning to develop novel, holistic solutions to real-world problems.

References

[1] **Alex Beutel**, Amr Ahmed, and Alexander J. Smola. ACCAMS: Additive Co-Clustering to Approximate Matrices Succinctly. In *Proceedings of the 24th International Conference on World Wide Web **(WWW)***, pages 119–129, 2015.

[2] **Alex Beutel**, Abhimanu Kumar, Evangelos E. Papalexakis, Partha Pratim Talukdar, Christos Faloutsos, and Eric P. Xing. FlexiFaCT: Scalable flexible factorization of coupled tensors on hadoop. In *Proceedings of the 2014 SIAM International Conference on Data Mining **(SDM)***, pages 109–117. SIAM, 2014.

[3] **Alex Beutel**, Kenton Murray, Christos Faloutsos, and Alexander J. Smola. CoBaFi: Collaborative Bayesian Filtering. In *Proceedings of the 23rd international conference on World wide web **(WWW)***, pages 97–108, 2014.

[4] **Alex Beutel**, B. Aditya Prakash, Roni Rosenfeld, and Christos Faloutsos. Interacting viruses in networks: can both survive? In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining **(KDD)***, pages 426–434. ACM, 2012.

[5] **Alex Beutel**, Markus Weimer, Tom Minka, Yordan Zaykov, and Vijay Narayanan. Elastic distributed bayesian collaborative filtering. In *NIPS workshop on Distributed Machine Learning and Matrix Computations*, 2014.

[6] **Alex Beutel**, Wanhong Xu, Venkatesan Guruswami, Christopher Palow, and Christos Faloutsos. CopyCatch: stopping group attacks by spotting lockstep behavior in social networks. In *Proceedings of the 22nd international conference on World Wide Web **(WWW)***, pages 119–130, 2013.

[7] Meng Jiang, **Alex Beutel**, Peng Cui, Christos Faloutsos, and Shiqiang Yang. A general suspiciousness metric for dense blocks in multimodal data. In *2015 IEEE International Conference on Data Mining **(ICDM)***. IEEE, 2015.

[8] Meng Jiang, Peng Cui, **Alex Beutel**, Christos Faloutsos, and Shiqiang Yang. Catchsync: catching synchronized behavior in large directed graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining **(KDD)***, pages 941–950. ACM, 2014.

[9] Meng Jiang, Peng Cui, **Alex Beutel**, Christos Faloutsos, and Shiqiang Yang. Catching synchronized behaviors in large networks: A graph mining approach. *ACM Transactions on Knowledge Discovery from Data **(TKDD)***, 2015.

[10] Meng Jiang, Peng Cui, **Alex Beutel**, Christos Faloutsos, and Shiqiang Yang. Inferring lockstep behavior from connectivity pattern in large graphs. *Knowledge and Information Systems **(KAIS)***, pages 1–30, 2015.

[11] Abhimanu Kumar, **Alex Beutel**, Qirong Ho, and Eric P. Xing. Fugue: Slow-worker-agnostic distributed learning for big models on big data. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics **(AISTATS)***, pages 531–539, 2014.

[12] Evangelos E. Papalexakis, **Alex Beutel**, and Peter Steenkiste. Network anomaly detection using co-clustering. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining **(ASONAM)***, pages 403–410. IEEE Computer Society, 2012.

[13] B. Aditya Prakash, **Alex Beutel**, Roni Rosenfeld, and Christos Faloutsos. Winner takes all: competing viruses or ideas on fair-play networks. In *Proceedings of the 21st international conference on World Wide Web **(WWW)***, pages 1037–1046. ACM, 2012.

[14] Neil Shah, **Alex Beutel**, Brian Gallagher, and Christos Faloutsos. Spotting suspicious link behavior with fBox: an adversarial perspective. In *2014 IEEE International Conference on Data Mining **(ICDM)***, pages 959–964. IEEE, 2014.

[15] Neil Shah, **Alex Beutel**, Bryan Hooi, Leman Akoglu, Stephan Gunnemann, Disha Makhija, Mohit Kumar, and Christos Faloutsos. EdgeCentric: Anomaly detection in edge-attributed networks. *arXiv preprint arXiv:1510.05544*, 2015.

[16] Chao-Yuan Wu, **Alex Beutel**, Amr Ahmed, and Alexander J Smola. Additive Co-Clustering of Gaussians and Poissons for Joint Modeling of Ratings and Reviews. *NIPS workshop on Nonparametric Methods for Large Scale Representation Learning*, 2015.