Bimal Viswanath Research Statement

Today's Internet-based web and mobile services are primarily *user-driven* (e.g., Facebook, Uber, Amazon, Airbnb). Users create, share, and endorse content that sustain these services. Inevitably, the security and robustness of these services are impacted by the behavior of its participants. Malicious actors with financial, or competitive incentives are injecting fake accounts (Sybil identities), or compromising existing users accounts (*e.g.,* via malware infection). These attackers then profit at the expense of normal users, resulting in their financial loss (*e.g.,* e-commerce fraud), identity theft, exposure to manipulated information, and participation in viral political propaganda. Service providers also suffer financially or through the loss of valuable users. Left unchecked, these attacks can lead to the collapse of some of the most popular online services today.

My research aims to understand and limit malicious activities in user-driven services. I take an interdisciplinary approach and play the dual role of a data scientist and a systems researcher to tackle security challenges. As a systems researcher, I use empirical measurement techniques to identify attacks targeting real Internet-scale systems, and build defenses that are practically deployable and satisfy system requirements, *e.g.,* scalable to large user base. However, these security challenges are hard to solve without incorporating data-driven algorithmic techniques because we need to analyze behavior of users in the system. User behavior tends to be complex, multi-dimensional, and hard to model. As a data scientist, I leverage a deep understanding of the strengths (and limitations) of different machine learning (ML) algorithms to model user behavior and characterize malicious activity. To develop defenses robust to intelligent adversaries, I design, test, and deploy algorithmic techniques based on behavioral data tied to resources that are hard for the attacker to forge or maintain.

The interdisciplinary nature of my work gave me the opportunity to publish in top tier venues in security (CCS, Usenix Security), systems (SIGCOMM, EuroSys, CoNEXT, IMC), and data mining (WWW, WSDM). My work received the Distinguished Paper Award at SOUPS'14 [8], and the Best Paper Award at COSN'15 [2]. My publications have been cited more than 2,700 times according to Google Scholar[1] (as of December 2017). My work also received popular press coverage by outlets such as Scientific American, Business Insider, Forbes, and New Republic. I frequently work with real world data about user and system behavior, which involves measurement effort to collect data at scale. Whenever possible, I release measurement data (in a privacy-preserving manner) to benefit the research community at large. For example, data I collected for a measurement study on user interaction on Facebook [3] has been used by over 570 research groups across the globe.

In the next few sections, I explain my previous and ongoing research work as well as future directions.

1 Resisting information manipulation using ML

User-driven sites are changing the way we consume information. Instead of relying only on traditional mass media (*e.g.,* CNN, NYTimes), many of us obtain news updates on social media sites (Facebook, Twitter), and read reviews/opinions about products/businesses on online marketplaces (*e.g.,* Yelp, Amazon). As users are given the power to generate and curate content, it also enables malicious entities to *manipulate the information we consume*. Examples include fake product reviews to mislead customers, fake news articles to sway public opinion, and social engineering attacks that lead to financial loss.

The growth of these attacks coincides with the rapid growth and maturation of machine learning systems, particularly deep learning neural networks, which are now capable of a variety of complex tasks previously considered impossible for software systems. A significant area of my research examines the potential consequences of information manipulation attacks using advanced machine learning tools.

For example, many online attacks today use *crowdturfing platforms*, where human workers are hired to write deceptive fake content targeting content-sharing sites. Fortunately, two factors limit such attacks, mainly, the cost of hiring human workers, and the predictable posting patterns of malicious human workers (detectable using ML techniques). An ML program that can create fake content would reduce the attack cost (humans are no longer required) and also provide full software control to evade existing ML-based defenses. Eliminating human effort also enables attacks at scale. Advances in deep neural networks (DNN) have reached a stage where such attacks are possible.

[1] http://scholar.google.com/citations?user=bu1SCOuD_CMC&hl=en

I led a team that took the first step towards evaluating such attacks, by considering platforms that use short form text. Our study focuses on online review systems such as Yelp, which are plagued by fake reviews. We show that an ML program based on *Recurrent Neural Networks (RNNs)* are capable of generating deceptive yet realistic looking reviews targeting restaurants on Yelp [11]. Note that generative language models like RNN are still not sophisticated enough to truly mimic human writing. However, our insight is that they are capable of mimicking short, domain-specific user generated content, *e.g.,* online reviews.

Extensive evaluation of the machine generated reviews shows they are undetectable using existing ML-based classifiers. We also conduct a user study, and find that not only do real users consistently fail to identify synthetic reviews, but they also perceive them to be as "useful" as real reviews written by other users. This attack is practical and highly effective on systems today, *e.g.,* Yelp, Amazon, Twitter. To protect these services, we developed a robust detection scheme that uses low level stylometric signatures to detect synthetic content. While the defenses are extremely effective today (95% detection rate with low false positives), their efficacy will degrade as larger neural networks are used for attacks.

I am working to extend this work to understand potential attacks that generate longer pieces of synthetic content, such as news or blog articles. One approach we are studying is the effectiveness of neural networks (DNNs) that "manipulate" the sentiment or message inside existing long-form content. Humans typically do not write complex text in a single pass, instead we perform multiple edit passes over the text to arrive at the final version. Inspired by this, I plan to study DNN based schemes that can learn to edit existing content to create new fake content, and potential defenses against them.

As DNN algorithms become more accessible with the emergence of Machine Learning as a Service platforms [12], it is important to understand how it can be used as an attack tool. My work highlights the need to prepare for artificial intelligence (AI) based attacks that can manipulate information we consume on the Internet.

2 Robust data-driven schemes to mitigate malicious activities

Much of my prior work investigated techniques that leverage user behavior to mitigate malicious activities. User behavior information can include any activity associated with a user (*e.g.,* liking a page on Facebook or messaging a friend), including user generated content.

Attacks typically exploit the privileges associated with a user account or identity. In most services, creating an identity is easy (*i.e.,* no strong verification) [10]. Therefore, they are vulnerable to *Sybil attacks*, where a large number of Sybil (fake) identities are created to abuse the system. We also observe cases of non-Sybil identities (*i.e.,* not fake) being incentivized to *collude* in order to manipulate each others' popularity, and also cases where non-Sybil identities are *compromised* by attackers to abuse services.

Enabling robust crowd computations. Our approach, called `Stamper` [2], is based on the idea that even when it is fundamentally hard to distinguish between individual Sybil and non-Sybil identities (*e.g.,* due to limited activity information), large *groups* of Sybil and non-Sybil identities can be differentiated. Looking at groups of identities makes sense because services are increasingly employing *crowd computing* to rate/rank content, users or businesses by polling the "wisdom" of the crowd (a group of users). For example, Yelp leverages crowd opinion to rate restaurants and it is useful for the provider to know if the crowd participating in a rating computation is malicious (or contains Sybil participants).

Our insight is that if an attacker tampers a computation using a large number of Sybil identities, it would result in an anomaly in the distribution of activity-levels of the crowd participants. We leverage activity features that attackers cannot forge, namely, the timestamps of their activities (*e.g.,* join date timestamps), thus raising the bar against an adaptive attacker. Using Stamper, we deployed real systems to detect users with manipulated follower counts, and content with manipulated popularity in Twitter. To further defend against repeated manipulation by identities, we also analyze individual user behavior (whenever available) for anomalous patterns [1]. Unlike most prior work, we detect a variety of malicious identities, including real-world Sybil, compromised and colluding identities. A highlight of this work was investigating the Facebook ad platform for evidence of *click-fraud*. A majority of clicks received for ads we ran, looked anomalous. This was one of the first studies to examine the click-fraud problem in a social ad platform.

Building Sybil tolerant systems. I proposed a novel approach called *Sybil tolerance* [4] to limit malicious activities driven by pairwise user interactions, *e.g.,* spam, fraud in buyer-seller transactions. Our idea is that instead of trying to detect Sybil identities, we can limit the impact that Sybil identities have on non-Sybil identities (*e.g.,* limit spam messages). A key assumption is the availability of an underlying social network between users (*e.g.,* formed from friendship relationships), and the attacker's inability to establish an arbitrary number of (friendship) links with real users [5, 9]. The impact that Sybil identities can have on non-Sybil identities is bounded proportional to the size of the min-cut separating them in the social network. We developed a generic library called `Canal` [4, 7] to make applications (with a social network) Sybil tolerant. Our implementation scales to large social networks with millions of users. Since then, other researchers have leveraged `Canal` to enable scalable transactions in real world payment settlement networks like Ripple.

3 Data-driven analysis of user privacy

My work in this space focuses on understanding the privacy vulnerabilities of popular social networking platforms and helping users better manage their privacy preferences. I conducted one of the first studies to investigate privacy leakage for implicit data in social networks [6]. I demonstrated that, even for data that is not explicitly shared by the user, the linked nature of information in social networks can lead to privacy leakage. In another work, we focused on user privacy related to content that is explicitly shared by the user [8]. We conducted the first large-scale study of user privacy specifications of over 1,000 users on Facebook and evaluated the challenges with building a tool to automatically recommend privacy preferences for a user. This work received the Distinguished Paper Award at SOUPS'14.

4 Future work

My future research agenda is influenced by the rapid progress made in developing new AI algorithms and related hardware. Advances in deep learning are significantly outperforming traditional ML systems in a variety of predictive and modeling tasks. As AI tools become commoditized, I will investigate two different scenarios: one where they can be used by bad actors as attack tools, and one where their vulnerabilities can be exploited by attackers to compromise new security systems, *e.g.,* payment systems using facial recognition for authentication. In each case, I plan to experimentally understand the capability and limitations of potential attacks, and develop robust defenses that help secure the adoption of ML-based systems.

4.1 Long term plan

Re-thinking data-driven security considering an AI-powered adversary. For years, the security community has made assumptions that attackers have limited algorithmic intelligence. But with the rise of AI and DNNs, that assumption needs to be reassessed. *What attacks are possible when bad actors make intelligent use of AI and neural networks?* An AI powered adversary can learn from data (*e.g.,* about system or user behavior), and render existing defenses ineffective. This would require us to completely re-think traditional data-driven security. I plan to investigate attacks that go beyond manipulating the information we consume (discussed earlier). I envision this to be a long-term plan, because both hardware and algorithmic capabilities are constantly evolving, and can hugely impact both attack and defense capabilities.

The examples are many. An AI powered malware package could be trained to automatically identify vulnerable hosts based on behavioral patterns, and self-organize to launch more devastating distributed attacks. AI could assist attackers in devising defense countermeasures by minimizing adaptation time and augmenting capabilities. Transfer learning techniques can be used to quickly tune an existing attack model to change its characteristics. In online profiling: deep learning techniques could enable network eavesdroppers to perform previously impossible de-anonymization attacks on encrypted or anonymized communication channels.

Another category of attacks includes those that *blur the boundary between human and machine-generated behavior*. AI could override user behavior based defenses by generating behavioral sequences mimicking real users (*e.g.,* a sequence of activities on a site). Deep generative models can be used for such attacks. This can enable an army of Sybil identities or fake virtual devices (when we consider mobile applications) that exhibit real behavioral traits, while trying to achieve malicious goals. Attacks can also impact the way we interact with other users on the Internet. Conversational AI models (*e.g.,* sequence to sequence models) could mimic content involved in interactive apps to automate large-scale social engineering attacks. The potential impact is significant, given

how many companies are moving to chat-based interfaces for customer services, and social chat platforms (*e.g.,* WeChat) are immensely popular as tools for mobile payments, and navigation.

Building robust defenses to these attacks is fundamentally challenging, because the attacker has (potentially) equally effective algorithmic tools. Hypothetically, whatever an operator chooses to use as a detection mechanism, an attacker can always tune their algorithmic approach to train for and evade it. Hence, we need to first understand the scope and limitations of AI powered attacks in different application scenarios. Note that deep learning models used today, are still limited by a finite learning capacity. In other words, assuming a reasonable computational budget, there is only so much information a model can learn from the training data. For example, a model that mimics human behavior may fail to capture all the complex characteristics of real user behavior. Identifying such limitations is non-trivial, but should help with defenses in different scenarios.

4.2 Short term plan

Enabling secure sharing and re-use of AI models. Developing DNN models require an enormous amount of training data and considerable expertise. Hence, the ability to build high quality models for complex tasks (*e.g.,* face recognition), typically lies at the hands of entities with such resources (*e.g.,* Google, Facebook). Fortunately, many of these trained models are being shared publicly on the web (*e.g.,* Inceptionv3), enabling others (with less resources) to build new AI-based applications, after applying some minor modifications to the models to suit their own tasks. While these trends make AI more accessible, they also raise certain security concerns. Any vulnerability or bug in a publicly shared pre-trained model can be exploited by an attacker to target all new systems that derive from those models. In fact, popular DNN models have been shown to have serious vulnerabilities in classification tasks when given certain adversarial inputs. Thus, the key security challenge is to make model re-use more secure and robust in an adversarial scenario, and to enable updates to derived systems when new vulnerabilities are discovered in the parent model.

Building ML-based defenses with limited labeled data. A practical hurdle with using ML to build user behavior based defenses is the limited availability of labeled data for both normal and malicious behavior. However, unlabeled behavioral data is plentiful. One idea is to leverage advances in deep learning that significantly improves semi-supervised learning using limited labeled data and a large amount of unlabeled data. These approaches usually assume some structure to the underlying data distribution. For example, training data may need to lie on a low dimensional manifold. While this may be the case for normal user behavior, it may not hold for malicious behavior samples. Thus, one challenge is to adapt these semi-supervised learning schemes for application in a security setting. Practical significance of this study would be huge, because it would allow emerging online services to step up their defense capabilities using limited labeled data.

References

[1] **Bimal Viswanath**, M. Ahmad Bashir, Mark Crovella, Saikat Guha, Krishna P. Gummadi, Balachander Krishnamurthy, and Alan Mislove. Towards Detecting Anomalous User Behavior in Online Social Networks. In *USENIX Security*, 2014.

[2] **Bimal Viswanath**, M. Ahmad Bashir, M. Bilal Zafar, Simon Bouget, Saikat Guha, Krishna P. Gummadi, Aniket Kate, and Alan Mislove. Strength in Numbers: Robust Tamper Detection in Crowd Computations. In *COSN*, 2015.

[3] **Bimal Viswanath**, Alan Mislove, M. Cha, and Krishna P. Gummadi. On the Evolution of User Interaction in Facebook. In *WOSN*, 2009.

[4] **Bimal Viswanath**, Mainack Mondal, Krishna P. Gummadi, Alan Mislove, and Ansley Post. Canal: Scaling Social Network-based Sybil Tolerance Schemes. In *EuroSys*, 2012.

[5] **Bimal Viswanath**, Ansley Post, Krishna P. Gummadi, and Alan Mislove. An Analysis of Social Network-based Sybil Defenses. In *SIGCOMM*, 2010.

[6] Alan Mislove, **Bimal Viswanath**, Krishna P. Gummadi, and Peter Druschel. You Are Who You Know: Inferring User Profiles in Online Social Networks. In *WSDM*, 2010.

[7] Mainack Mondal, **Bimal Viswanath**, Allen Clement, Peter Druschel, Krishna P. Gummadi, Alan Mislove, and Ansley Post. Defending Against Large-scale Crawls in Online Social Networks. In *CoNEXT*, 2012.

[8] Mainack Mondal, Yabing Liu, **Bimal Viswanath**, Krishna P. Gummadi, and Alan Mislove. Understanding and Specifying Social Access Control Lists. In *SOUPS*, 2014.

[9] Saptarshi Ghosh (co-primary), **Bimal Viswanath (co-primary)**, Farshad Kooti, N. K. Sharma, Gautam Korlam, Fabrício Benevenuto, Niloy Ganguly, and Krishna P. Gummadi. Understanding and Combating Link Farming in the Twitter Social Network. In *WWW*, 2012.

[10] Giridhari Venkatadri, Oana Goga, Changtao Zhong, **Bimal Viswanath**, Krishna P. Gummadi, and Nishanth Sastry. Strengthening Weak Identities Through Inter-Domain Trust Transfer. In *WWW*, 2016.

[11] Yuanshun Yao, **Bimal Viswanath**, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. Automated Crowdturfing Attacks and Defenses in Online Review Systems. In *CCS*, 2017.

[12] Yuanshun Yao, Zhujun Xiao, Bolun Wang, **Bimal Viswanath**, Haitao Zheng, and Ben Y. Zhao. Complexity vs. Performance: Empirical Analysis of Machine Learning as a Service . In *IMC*, 2017.