Research Proposal

Jooyeon Kim

May 28, 2018

Throughout my Ph.D. studies, I aim to **(1) understand properties of individuals or groups** that are not directly observable but become more salient throughout their social interactions and **(2) design machine learning models** that can benefit each individual or the system as a whole. Individual characteristics and their collective behaviors are becoming more diversified with an increase of user interactions prevalent in online media, communities, social networking services, commercial sites, and games. We can benefit from learning such characteristics to develop strategies that can be learned from and automatically adjusted to the given environment. However, unveiling insightful properties of individuals or groups and designing new models require comprehension on existing machine learning models and their respective inference algorithms and careful reasonings from setting up the research hypotheses to drawing conclusions. My past and current research tackle such issues in different angles as follows:

(i) **Temporal Dynamics Modeling and Control**: I propose an optimal policy for a system or users by modeling (i) the temporal dynamics of the policy that I aim to optimize, (ii) the surrounding entities that comprise the system of interest, and (iii) the target variable which is controlled with respect to the policy and the surrounding entities. Next, I represent the temporal dynamics in the form of stochastic differential equations (SDEs) with jumps. Finally, by defining a loss function with the obtained SDEs and minimizing the expected value of the loss function with respect to the policy over the possible changes of the state variables, the optimal policy is obtained for any given time, letting us come up with an online algorithm for the decision-making policy.

(ii) **Probabilistic Machine Learning**: I design application-level probabilistic machine learning models and propose approximate inference algorithms or sampling methods for the models. To unveil interesting user properties, I exploit different kinds of data such as text, network, time, and source identities. I represent each property as a random variable and design a probabilistic model that allows the variables to effectively interplay, exploits prior beliefs, and produces desired outcomes. I propose inference algorithms for the model posterior, which is usually intractable to calculate for complex models, by referencing existing statistical methods.

(iii) **Computational Social Science**: I construct novel research hypotheses for a given task with the aid of insights from existing social science research. I validate the hypotheses through large-scale data analyses and statistical testings. Aside from the data analyses, I conduct user interviews and surveys to qualitatively manifest the validity of the hypotheses.

1 Past and Current Research

During my graduate studies, I have published three papers in international conferences and a journal in the fields of data mining and information retrieval (WSDM), natural language processing (TACL), and human-computer interaction (CHI).

Optimal Policy for Suppressing Fake News Propagation. In recent years, social media and online social networking sites have served as major disseminators of false facts, urban legends, fake news, or, more generally, misinformation. I propose an optimal policy for detecting fake news and suppressing the propagation of misinformation using signals from the crowd [14]. I model temporal dynamics of the users and exploit user feedbacks such as flagging or reporting. Here, the problem is when to fact-check posts given a *sufficient* number of flags. The tradeoff is that when we report the article too early based on small samples, there is a high risk of the post being authentic (not containing misinformation), whereas if we report the post too late, a lot of people will be affected by misinformation if the post turns out to be fake. CURB is a scalable online algorithm that effectively negotiates the tradeoff by detecting suspicious posts and reducing the spread of misinformation. I use marked temporal point processes to represent the dynamics of users, target variables, and the policy and optimize the policy using the stochastic optimal control framework. I test the efficacy of the algorithm using two social network datasets, Twitter and Weibo. In Figure 1, CURB fact-checks suspicious posts before they go viral and suppresses misinformation spread better than the baselines.

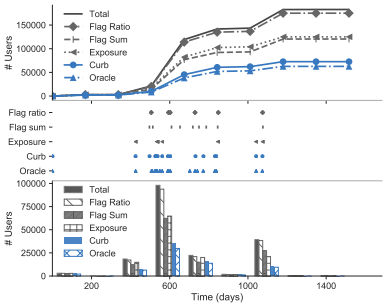

Figure 1: Misinformation reduction vs. time. The proposed control signal CURB prevents the spread of misinformation before it becomes viral.

Authoritative Authors in Citation Network. As a graduate student who just kicked off her graduate studies in machine learning, whose paper, among myriads of researchers, should she read first as a starting point? I propose a probabilistic model that gives answers to such questions by measuring authors' *topical* scholarly authority [13]. The proposed model jointly models academic papers' text, authorship, and the citation information and generates topics and authors' scholarly authority scores (Figure 2). I develop the model based on the assumption that a citation is more likely to occur (i) when two papers are topically similar, and (ii) when the topical authority of the cited paper's authors are high. I propose efficient inference method based on stochastic variational inference [10] to optimize the complex model with entangled random variables. Because of its joint-modeling nature, the content-related variables and the citation-related variables mutually reshape one another during the posterior inference. Finally, by incorporating the topical authorities the model predicts hidden citation and authorships better than previous models.

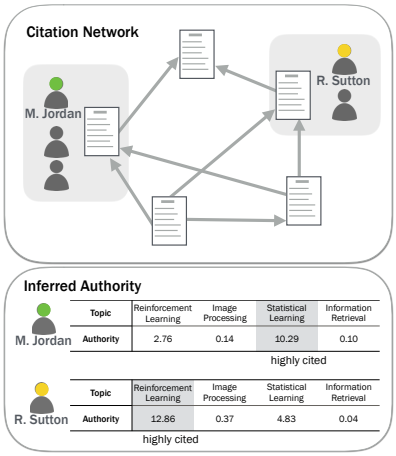

Figure 2: Overview of Latent Topical Authority Indexing. Based on content, citation, and authorship (top), LTAI discovers the topical authority of authors; it increases when a paper with certain topics gets cited (bottom).

Individual and Team Strategies in Games. Suppose that five people who have never met before are randomly assigned to a team. They are instructed to pick a role and work together to accomplish a certain task while competing with another group of the same size. In these days, gamers who play multiplayer online battle arena (MOBA) games encounter situations like this on a daily basis. My research on individual and team strategies in games [12] analyzes the compositional traits of over a million teams in a famous online game called League of Legends. In the research I hypothesize that gamers must negotiate a *proficiency-congruency dilemma* between selecting (i) roles that best match their experience and (ii) roles that best complement the existing roles on the team and increase its diversity when they assemble their teams and try to maximize their performance.

To confirm or reject the hypothesis, I first map LoL characters in the champion similarity space, where the mapping is based on their abilities and roles, and cluster the characters in 5 non-overlapping groups (Figure 3). Then, I quantitatively define user proficiency and team congruency by calculating the distance between the characters. Using the two metrics that represent two distinctive compositional strategies, I confirm the hypothesis by showing that the proficiency and the congruency negatively correlates one another. Remarkably, however, the negative correlation gradually diminishes as we go from underachieving teams to elite teams, hinting that successful teams can finesse the dilemma and maximize the team performance better than the novice groups. Finally, I fortify the confirmation of the hypothesis by conducting focus group interviews with both elite and novice game players.

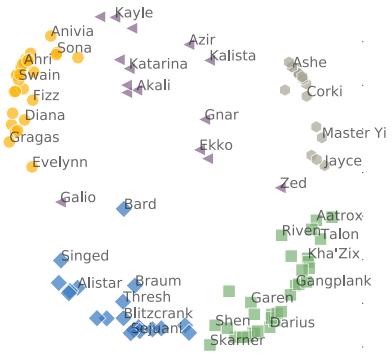

Figure 3: Champion functional clusters. I use PCA to reduce the champion feature space to the most salient features, and use k-means to partition the champions into five clusters.

2 Future Direction

Stemming from the above research trajectory, my long-term research goal is to **frame machine learning models that can mimic human decision-making processes based on the temporal, textual, and social traces of online users**. When we dichotomize the intelligence into compute information vs. communicate information, humans excel in the computing realm, *i.e.*, abstract the system that surrounds us so that we can comprehend it, while machines are better at communication, *i.e.*, efficiently read a vast amount of data and display the outcome [15]. Can machines model a complex system without overlooking diverse kinds of data? Can machines learn a system even when the full knowledge of it is not given? I believe that by trying to answer these questions and by making machines perceive and abstract the system the way humans do, they can benefit us in a much better fashion than now, fully exploiting both sides of the intellectual realms.

In the next couple of years, I envision to take small steps towards my goal in different dimensions as follows:

Model. I will design novel machine learning models with focuses on the following aspects:

1. **Incorporating homogeneity for true and fake news modeling in social networks:** In recent years, users in social networking services and microblogging platforms have played major roles in disseminating online content and news stories [24]. While some are based on factual information, a non-negligible portion of the news stories contains false information that is used maliciously to stir unnecessary disputes, sway people's opinion, and give rise to polarization in the community [3, 18]. However, tracking the diffusion pattern of both true and fake news stories requires a holistic comprehension of a multi-faceted system that comprises of multiple submodules such as: content and source information of the news stories, interests of users and their topical alignments, and social connections among the users [24]. Also, Vicario et al. have suggested *homogeneity* as the key driving force for the diffusion of news stories in online social networks [6]. That is, for the similar virality, some stories will spread across the whole social network, while others will propagate within a confined portion of the network. In my Ph.D. studies, I will focus on understanding the interplay of the news homogeneity, genuineness, content, and users' topical interests by formulating a Bayesian nonparametric probabilistic model. In the modeling aspect, I focus on recent advances in nonparametric topic modeling [11, 20, 22] to model contents and the topical transmission from one user to another and Gaussian process latent variable model (GP-LVM) [23], which accelerate the inference speed and accommodate multidimensional outputs and latent inputs in highly nonlinear fashion, to model the homogeneity of news stories and its genuineness. The model can be complementarily implemented with other social network and information propagation-related probabilistic models [2, 7, 9, 17] to comprehensively understand the interplay among diffusion patterns, homogeneity and user interests as well as their social connections.

2. **Leveraging different kinds of data:** Events in online social networks usually entail information such as time, texts, and generators' identities. Hawkes process [8] is a branch of point processes and is useful in modeling the real-world event generation patterns by uncovering the triggering/triggered relationships among events and quantifying influence of users that generate the events. In the upcoming months, I will define variants of Hawkes processes that operate on *multidimensional* input spaces. While there are previous works that incorporate textual information in Hawkes processes [7, 9, 17] and point processes that operate with vectorized inputs [1, 19], a connection between the general multidimensional input spaces and Hawkes processes has not been explored. The preliminary results drawn by the model shows promising results for both expanding its applicability for various inputs and predicting future events.

3. **Modeling uncertainty:** Temporal point processes lack the ability to model uncertainty. For example, following the immigration-birth representation of the Hawkes process, parents and children have a one-to-one, deterministic relationships. On the other hand, Gaussian processes are highly affective at generating outcomes with uncertainty and there has been recent advances in the field that allows modeling of dynamical patterns in sequential data [16]. I believe that it is possible to jointly model recurrent Gaussian process with Hawkes process to have expected outcomes of uncovering long-term dependencies and stochastic relationships in event cascades and information diffusions.

Algorithm. I will be attentive to the advances in the state-of-the-art inference algorithms for different kinds of models so that I can readily go through the inference processes for novel model assumptions of my own. I am particularly interested in recent breakthroughs in Gaussian processes,

not only in the models, but also in the way researchers apply and develop different inference algorithms to tackle different issues such as model generalization, and algorithm speed-up and relaxing unrealistic assumptions within [4, 5, 21].

Data. I will constantly seek for new datasets that encompass novel and multifaceted interaction patterns of users and give rise to challenging machine learning tasks. I will explore crowd-learning communities such as Stack Overflow, content rating and discussion websites such as Reddit, and online encyclopedias such as Wikipedia. While these datasets are readily accessible online, I also plan on doing internships at IT companies and research labs where I can investigate on datasets with both industrial and academic focuses.

References

[1] Ryan Adams, Iain Murray, and David MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the Annual International Conference on Machine Learning (ICML)*, 2009.

[2] Amr Ahmed and Eric Xing. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the Association for Uncertainty in Artificial Intelligence (UAI)*, 2010.

[3] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one*, 10(2):e0118093, 2015.

[4] Zhenwen Dai, Andreas Damianou, Javier González, and Neil Lawrence. Variational auto-encoded deep Gaussian processes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[5] Andreas Damianou, Michalis Titsias, and Neil Lawrence. Variational inference for latent variables and uncertain inputs in Gaussian processes. *The Journal of Machine Learning Research (JMLR)*, 17(1):1425–1486, 2016.

[6] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 113(3):554–559, 2016.

[7] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander Smola, and Le Song. Dirichlet-Hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2015.

[8] Alan Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[9] Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor, and Yan Liu. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

[10] Matthew Hoffman, David Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research (JMLR)*, 14(1):1303–1347, 2013.

[11] Dongwoo Kim and Alice Oh. Hierarchical Dirichlet scaling process. *Machine Learning*, 3(106):387–418, 2017.

[12] Jooyeon Kim, Brian Keegan, Sungjoon Park, and Alice Oh. The proficiency-congruency dilemma: Virtual team design and performance in multiplayer online games. In *Proceedings of the Annual ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.

[13] Jooyeon Kim, Dongwoo Kim, and Alice Oh. Joint modeling of topics, citations, and topical authority in academic corpora. *Transactions of the Association for Computational Linguistics (TACL)*, 5(1):191–204, 2017.

[14] Jooyeon Kim, Behzad Tabibian, Alice Oh, Bernhard Schoelkopf, and Manuel Gomez-Rodriguez. Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2018.

[15] Neil Lawrence. Living together: Mind and machine intelligence. *arXiv*, 2017.

[16] César Mattos, Zhenwen Dai, Andreas Damianou, Jeremy Forth, Guilherme Barreto, and Neil Lawrence. Recurrent Gaussian processes. In *Proceedings the International Conference on Learning Representations (ICLR)*, 2016.

[17] Charalampos Mavroforakis, Isabel Valera, and Manuel Gomez-Rodriguez. Modeling the dynamics of online learning activity. 2016.

[18] Delia Mocanu, Luca Rossi, Qian Zhang, Marton Karsai, and Walter Quattrociocchi. Collective attention in the age of (mis) information. *Computers in Human Behavior*, 51:1198–1204, 2015.

[19] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

[20] John Paisley, Chong Wang, and David Blei. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(4):997–1034, 2012.

[21] Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2017.

[22] Yee Whye Teh, Michael Jordan, Matthew Beal, and David Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2005.

[23] Michalis Titsias and Neil Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[24] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.