# Building a heterogeneous social network recommendation system

Parag Agrawal October 6, 2020



*Co-authors: Parag Agrawal, Ankan Saha, Yafei Wang, Aastha Nigam, and Eric Lawrence*

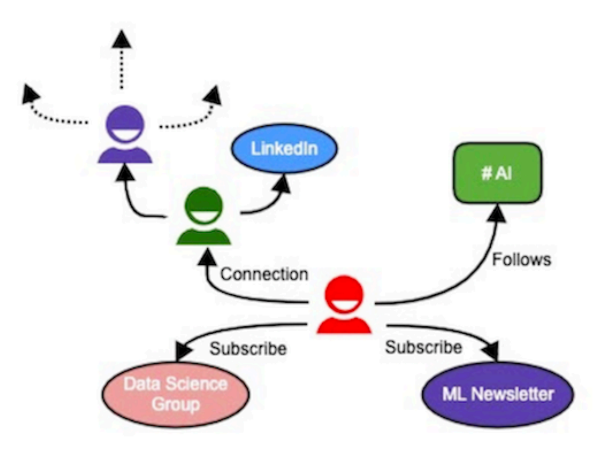


*Figure 1: A heterogeneous social network*

LinkedIn's "People You May Know" (PYMK) feature has long been used by our members to form connections with other members and expand their networks. As member networks become more heterogeneous, the People You May Know tab (MyNetwork tab hereafter) has evolved to show people, hashtag, company, group, newsletter, and event recommendations. When members act on these recommendations, they are adding edges to the graph that constitutes their social network. They can connect to another member (a "connection edge"); follow a hashtag, company, or influential-creator (a "follow edge"); and subscribe to a newsletter, group, or event (a "subscribe edge"). As members have increasingly turned to LinkedIn not just to find people, but also to build community and keep up to date on professional news, these edges can help members get access to relevant content and form active communities with which they can regularly engage.

## What makes heterogeneous edges different

These heterogeneous edge types are distinct in character and serve different value propositions. A connection edge is two-way or bidirectional, allowing both the inviter as well as the receiver of the invitation to have access to each other's content after edge formation. On the other hand, follow and subscribe edges are unidirectional, giving following and subscribing members access to content from other companies, hashtags, creators, groups, newsletter, and events. Recommendations in the My Network tab help members build a heterogeneous social network consisting of a) *heterogeneous edges*, namely connection, follow, and subscribe and, b) *heterogeneous entities*, namely people, hashtags, companies, groups, newsletters, and events.

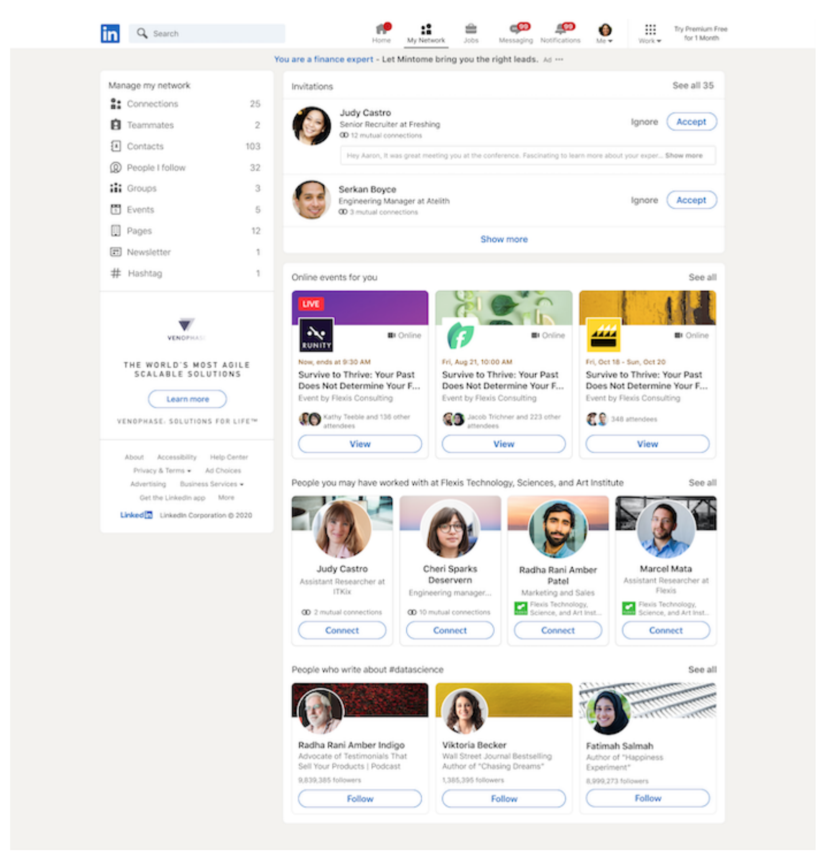


*Figure 2: The "My Network" tab showing heterogeneous recommendations "People you may have worked with" is a cohort of people recommendations that a viewer can create connection-edges with, while "People who write about #datascience" is a cohort of recommendations of influential-creators that a viewer can create follow-edges to.*

**Two-phased edge recommendations**
Given the heterogeneity of the diverse entities, ranking these recommendations typically follow a two-phase process:

1. Ranking entities of one type among each other: As depicted in the image above, this would entail ranking the people entities, "Judy," "Cheri," "Radha," and "Marcel," among themselves. This typically happens via preliminary rankers called Edge-FPR (Edge First Pass Ranker) models, and the entities are grouped together into a cohort of recommendations and presented to the members.
2. Ranking heterogeneous cohorts of entities *against* each other: For example, ranking a cohort of events vs. a cohort of people recommendations from your company vs. a cohort of newsletters. This facilitates the process of a member selecting *the next edge* to grow his or her heterogeneous network. To this end, we built a Second-Pass-Ranking (SPR) recommendation system.

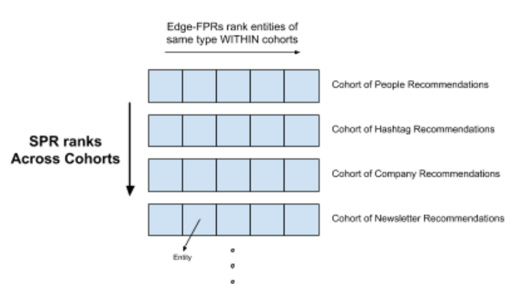


*Figure 3: SPR ranks across heterogeneous cohorts*

Developing such a recommendation system comes with three primary challenges:

1. The Edge-FPR models are separate and independent AI models that can leverage different algorithms such as XGBoost, logistic regression, and deep neural networks. These models harness the domain knowledge as a result of which they might optimize for different business metrics and might even generate scores on a different scale. Therefore, in order to generate a consistent ranking across the heterogeneous entities, we need to calibrate the Edge-FPR model scores to make them comparable.
2. Distribution of training data might not be indicative of the relative value of the different types of edges and could significantly diverge from the ideal distribution. Moreover, we need to take different business needs and limitations into account while ranking. Consequently, a uniform ranking of edges (that simply ignores the true relative value of edges) at the second pass scoring level could end up giving undeserved higher (or lower) importance to certain edge types.
3. Establishing the contract between Edge-FPR models and SPR is non-trivial, and requires a careful tradeoff between modeling accuracy and development agility. To improve the overall system, Edge-FPR models and SPR need to be regularly iterated upon through A/B experimentation. However, it's important to note that, while a tight coupling between Edge-FPR models and SPR could yield slightly better accuracy, it may slow down the iterations velocity.

Next, we will discuss the SPR framework and our strategy to address the aforementioned challenges. Before we dive into the details of the SPR model, let's start by defining some terms:

| | |
|---|---|
| **Entity** | An item or individual recommendation shown to the member. |
| **Cohort** | Grouping of entities of the same type which is then shown as a horizontal carousel on the MyNetwork tab. |
| **Entity types** | PYMK, Hashtag, Company, Member-Follow, Groups, Newsletters, Events |
| **Edge types** | CONNECTION edge: A member connects to another member.<br>FOLLOW edge: A member follows a hashtag, company, or member.<br>SUBSCRIBE edge: A member subscribes to a group, newsletter, or event. |

## SPR algorithm in brief

We develop an XGBoost model that predicts the probability of *downstream-interactions* of a member with top-*k* entities (entities occupying first *k* positions) within a cohort and ranks the cohorts using this probability score. A like, comment, or a re-share on content produced by the entity is counted as a downstream interaction; so for a connection-edge, this would mean number of likes, comments, or re-shares on the content posted by the connection. This model trains against a logistic loss with binary labels (corresponding to if there was a downstream-interaction or not) and uses *calibrated* scores from Edge-FPRs as features in addition to other member-level features. Further, we also design *counterfactual experiments* to estimate the relative importance of each edge type in the form of *importance* factors that are multiplied to the scores of the corresponding cohorts.

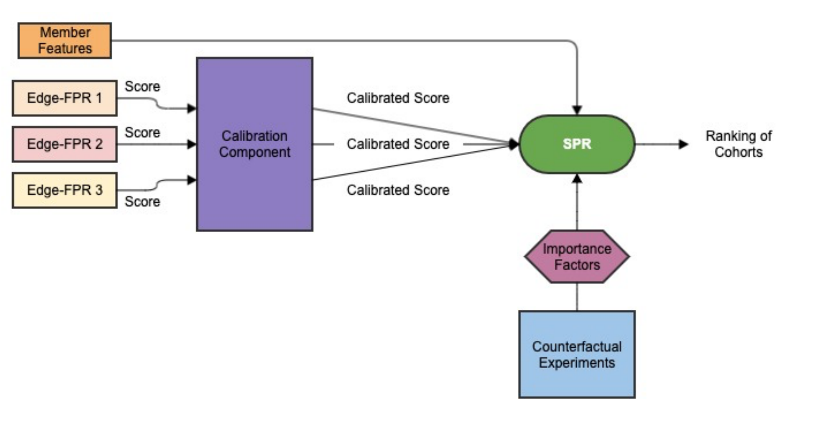


*Figure 4: Schematic design of SPR*
*a) Calibration component takes the raw Edge-FPR scores and outputs the calibrated scores. b) An importance factor for each edge-type is estimated from the counterfactual experiments. c) SPR uses calibrated scores, importance factors and member features to generate a probabilistic score which is used to rank cohorts.*

**Model equation**

$$P(like\ or\ comment\ or\ share) = g(Aggregation(Calibrated\ Edge\ FPR\ scores), Member\ features)$$

We predict the probability (P) of downstream-interaction using aggregated and calibrated FPR scores as well as member features. Sample member features include profile-based features, activity features, and neural-network embeddings that capture the member's network topology. We use an aggregation function (could be a simple weighted mean aggregation or a complicated non-linear aggregation) to convert Edge-FPR scores to the cohort-level.

**Challenge #1: Calibration**
To make the scores from different Edge-FPR models comparable, we define calibration to be a function mapping from the quantiles of each Edge-FPR score to the discretized observed response (response that is used for training the Edge-FPR model and not our SPR). Using $s$ to denote the quartiles of score generated by an Edge-FPR model and, $z$ to be the response variable for this Edge-FPR, $f$ becomes our calibration mapping function as $z \sim fs$. The structure of $f$ is dictated by the empirical relationship between Edge-FPR score $s$ and the response variable $z$. For instance, $f$ could take a log-linear form over different transformations of $s$. Ultimately, the estimate of $f$ is used as the calibrated Edge-FPR score in our model equation.

**Challenge #2: Disproportionate behavior**
Member behavior can vary drastically across edge-types. Typically, members tend to *connect* to only a limited number of members, whereas they like to *follow* numerous hashtags to consume content related to it. Consequently, this might lead to overrepresentation of *follow* edges in our training data and disproportionately high scores for them. In an extreme case, a *follow* edge can always result as the default top suggestion for the next heterogeneous edge, even when creating a connection edge (connecting to a person) could be more valuable than a follow edge (following a hashtag or company).

To address this, we estimate the true value of a particular edge by leveraging counterfactual experiments, where we temporarily remove a portion of member's network (more precisely, we drop content over verticals, such as feed and notifications, by a filtering strategy for certain existing edges from the member's heterogeneous network) and observe the impact to their engagement (sessions activity, visits, etc.). The data from these experiments are used to estimate the *importance factor* for each edge-type that is multiplied to the scores of the corresponding cohorts.

**Challenge #3: Coupling between SPR and Edge-FPRs**


*Figure 5: Tight coupling between SPR and Edge-FPR entails multiple back-and-forth updates*
*a) A marginal improvement to an Edge-FPR model triggers a retraining of SPR as the Edge-FPR score has changed. b) An improvement to SPR also triggers a retraining of each of Edge-FPR models.*

We keep Edge-FPR and SPR completely independent to avoid back-and-forth updates that would cause slowness in experiment velocity. Mandating a retraining of SPR every time an Edge-FPR model is updated would slow down the iteration speed for FPR models. We consciously make a decision to prioritize the agility of the engineering teams developing Edge-FPRs and resort to obtaining the benefit from SPR retraining at its own cadence.

**Model training and random bucket**
For training our SPR, we maintain a holdout of 5% (random bucket) where we show a complete random shuffling of cohorts. Tracking data from this random bucket is used for training the SPR model. To avoid bias from cohorts ranking in Edge-FPR models, the same data could also be used for training Edge-FPR models by the individual responsible teams.

**Impression guarantees**
When introducing a new entity, our SPR model could end up undeservedly pushing cohorts of these entities down in the ranking, leading to limited exposure and training data for the teams to build their own FPR models. To avoid this, we provide some kind of stochastic impression guarantee for cohorts of each type. One way of providing a simplistic notion of impression guarantee is to get a read of the importance of the different edge types (connection, follow, subscribe), define a constraint on the number of cohorts that should be provided for that particular category, and use that as a guarantee to begin with. *Note that the system still remains flexible in terms of ranking these cohorts in any possible order.* For instance, a 1:2 importance for connection vs. follow edge, would entail ensuring twice as many impressions of follow cohorts to connection cohorts, while the ranking among them is dictated by the SPR.

Periodically, we would get fresher reads of this relative importance (based on iterations of the counterfactual experiments or how the SPR system performs in terms of metrics), and accordingly change the cohort impression guarantees on a periodic basis (monthly or quarterly). We specifically choose the guarantee of impressions at a *per-viewer and per-edge-type level*. This doesn't provide a global guarantee for each cohort. A global impression guarantee for *each cohort*—while ideally preferable—would be far too complex to operationalize for what it's worth.

## Results and next steps

To measure the impact of our SPR system, we conducted A/B tests, which showed an increase in the number of engaged members and a significant increase in the downstream-interactions of members. The new system helped **more** members not only create edges (e.g., connecting to other members, following hashtags, subscribing newsletters), but also have **conversations** over these newly formed edges.

An effect we see in our heterogeneous social network recommendation system is *cannibalization* across edge types. Formation of certain edges can come at the cost of other edges; while there might be an overall increase in the number of edges and members interactions, the distribution of this increase over the different edge types depends on the specifics of the SPR algorithm, chosen to appropriately satisfy the product specifications for *each* edge type. There is also *heterogeneity* in interactions among member groups. Frequent members continuously provide us with rich data to show high-quality recommendations, while inactivity from infrequent members leads to lack of data and lower-quality recommendations. We plan to address these limitations and continue to invest into our strategy of building more holistic and active communities on LinkedIn that help make all of our members more productive and successful.

## Acknowledgements

It takes a lot of talent and dedication to build the AI products that drive our mission of building active communities on LinkedIn. We would like to thank Aastha Jain, Yan Wang, Ashwin Murthy, Abdul Al-Qawasmeh, Albert Cui, Jugpreet Talwar, Zhiyuan Xu, Bohong Zhao, Judy Wang, Chiachi Lo, Mingjia Liu, David Sung, Qiannan Yin, Quan Wang, Jenny Wu, Andrew Yu, Shaunak Chatterjee, and Hema Raghavan for their instrumental support, and Shaunak Chatterjee, Yiou Xiao, Kinjal Basu, Michael Kehoe, Heyun Jeong, Stephen Lynch, and Jaren Anderson for helping us improve the quality of this post. Finally, we are grateful to our fellow team members from PYMK AI, Growth Eng, Communities AI, Optimus AI, and Growth Data Science teams for the great collaboration.

Topics

artificial intelligence, Recommender Systems, machine learning, Data