

# Graph Recommendations: Random Matchings on Graphs

Arda Antikacioglu, R. Ravi, Srinath Sridhar

June 26, 2013

## Abstract

In this paper, we study the problem of graph recommendations as variants of bipartite matching problems. We consider the problem of solving such matching problems in practice at web-scale. To achieve this we introduce several models to simulate underlying input graph structures. We then analyze the conditions under which a random sample of edges using constant memory already suffices to be a 'good' recommendation algorithm as opposed to the cases when we may consider the more classical and involved linear memory polynomial time algorithms. We also show how to select the number of recommendations per item while building a website so that there exists a 'perfect' graph recommendation.

## 1 Introduction

Recommendations are an integral part of several popular websites. Specifically websites that are highly engaging and typical user behavior involves that of long browsing sessions almost always contain a recommendation system. For e.g., YouTube recommends a set of videos on the right for every video watched by a user, Quora recommends a set of questions on every question page, Amazon recommends similar products on every product page etc. For the rest of this Section, we will use YouTube as a running example although these observations apply to all websites and recommendation systems.

Note that these recommendations are so critical that one could argue that the YouTube would not be the website it is today if it had no recommendations. A seamless browse experience that transitions the user from one YouTube video to another directly dictates the amount of time a user spends on the site and is therefore directly correlated to both user satisfaction and revenue.

While recommendations are important, its implementation is performed by finding related items at runtime. That is, when a user lands on an item the recommendation systems decide online the set of other items to display. Such systems have the problem that a global analysis of the performance of the recommendations are hard. For e.g., one can simply ask if there are an item that was not recommended by even one other item?

This motivates solving the recommendation problem offline where we can precompute the set of recommendations for every item. Such a system can be built in two stages. A first stage decides on recommendations for each item purely based on relevance and retrieves  $d$  items to show. The second stages prunes it to  $c < d$  items such that globally we optimize to ensure that the resulting recommendation graph is 'good'. For e.g., we might want to ensure that the resulting graph minimizes the number of items that was not recommended by a single other video.

We can represent this notion of recommendations by using a directed graph. A vertex is simply an item and a directed edge  $(u, v)$  is a recommendation from  $u$  to  $v$ . Under this graph recommendation,

for the above problem, if  $c = 1$  and we require the graph be strongly connected then our problem reduces to Hamilton cycle, an NP-complete problem.

We can generalize the above graph recommendation problem by using a directed bipartite graph. Website owners might not be interested on all the items on their site. We can use one partition  $U$  to represent the set of items for which we need recommendations. We can use the other partition  $V$  to represent the set of items that are recommended. Note that a single item can be represented in both  $U$  and  $V$  if needed. For our main results we will use this representation. Now the input to the problem is this directed graph  $G$  where each vertex has  $d$  recommendations (thanks to the first stage of the above described backend system). The output that we need is a subgraph  $G'$  where each vertex has  $c < d$  recommendations. The goal is simply to minimize the number of vertices that have in-degree less than  $a$ . Again, note that if  $a = c = 1$  this is simply the problem of maximum bipartite matching [1]. If  $a = 1$  and  $c > 1$ , this is the problem of fractional matching [2]. Both these problems have well-studied polynomial time algorithms [3]. For the rest of the paper, we call this the  $(c, a)$ -graph recommendation problem.

In practice, the authors have implemented several algorithms in cutting-edge web-technology companies including Google, Facebook and BloomReach. There are a couple of massive hurdles in the practicality of backend systems. First is just simplicity. Software and systems need to be simple for it to be maintainable. It is a nightmare to maintain a complicated system and is avoided almost always in practice unless absolutely needed. Second is scale. Systems need to scale gracefully and easily. Matching algorithms require linear memory and super-linear run-time neither of which scales. Facebook graph has a billion vertices [4] and hundreds of billions of edges [5], YouTube has XXX videos and YYY recommendations [6] etc. Even an ecommerce website with 100M product pages and 10 recommendations per product would require 100+GB in main memory to run these algorithms. Since using clusters of single machines with such high memory is prohibitively expensive, map-reduce [7] jobs are needed to solve these problems. Map-reduce for graph problems are firstly notoriously hard and secondly we are back into building a complicated system.

The above reasons motivated the authors to study the problem of recommendations as a graph matching problem and analyze the cases when a simple set of recommendations would suffice and produce near optimal solutions. Our goal is to find simple relationships between the parameters of the problem and compute thresholds that dictate the need for the different algorithms. In practice, this would immediately imply that a very simple backend system can be built in many cases without the need for a complicated algorithm. If the thresholds provided by our analysis are insufficient in quality, then the designer can consider implementing the classical algorithms.

## 2 Summary of Results

Our main contributions are the following:

1. We describe several general models such as the Hierarchical Tree Model and the Cartesian Product Model which describe how graph recommendation subgraphs arise probabilistically.
2. For each of these models, we describe a simple randomized linear time algorithm that solves the graph recommendation problem to near optimality in expectation.
3. We also show that a simple linear time greedy algorithm can approximate the problem to a constant factor in the worst case.

4. We prove that optimal recommendation subgraphs that satisfy the most trivial upper bounds exist with high probability even in random bipartite graphs with logarithmic average degree.
5. We show that the models described in the problem can describe real-life recommendation subgraphs and that the random sampling algorithms we describe find near optimal solutions to the graph recommendation problem.

## 3 Results

### 3.1 Fixed Degree Model

In this model, we assume that a bipartite graph  $G = (L, R, E)$  is generated probabilistically by the following procedure. Each vertex  $v \in L$ , uniformly samples a set of  $d$  neighbors from  $R$ . For convenience let  $|L| = l$ ,  $|R| = r$  and  $k = l/r$ . From  $G$ , we sample a subgraph  $H$  where for each vertex  $u \in L$  a set of  $c$  neighbors are sampled uniformly from set of incident vertices. The following theorem derives a lower bound on the expected size  $S$  of the number  $v \in R$  such that  $\deg_H(v) \geq 1$ .

**Theorem 1.** *Suppose that  $G = (L, R, E)$  and  $H \subseteq G$  is generated as above. Then*

$$E[S] \geq r(1 - \exp(-ck))$$

where the expectation is over the random sampling of  $G$  and  $H$ .

*Proof.* For each  $v \in R$  let  $X_v$  be the indicator variable for the event that  $\deg_H(v) \geq 1$ . Note that since for each vertex  $u \in L$ ,  $H$  uniformly samples from a uniformly sampled set of neighbors, we can think  $H$  as being generated by the same process that generated  $G$ , but with  $d$  replaced with  $c$ . Now for a specific vertex  $u \in R$ , the probability that it has no incident edges is  $(1 - \frac{1}{r})^c$ . Since the selection of neighbors for each vertex in  $L$  is independent, it follows that that:

$$\Pr[X_v = 0] = \left(1 - \frac{1}{r}\right)^{cl} \leq \exp\left(-c \cdot \frac{l}{r}\right) = \exp(-ck)$$

Note that  $S = \sum_{v \in R} X_v$ . Applying linearity of expectation, we get

$$E[S] = \sum_{v \in R} E[X_v] \geq r(1 - \exp(-ck))$$

□

While this shows a lower bound in absolute terms, we must compare it to the best possible solution  $OPT$ . The follow theorem proves the approximation ratio to  $OPT$ .

**Theorem 2.** *The above sampling algorithm gives a  $1 - 1/e$  factor approximation to the  $(c, 1)$ -graph recommendation problem in expectation*

*Proof.* The size of the optimal solution is bounded above by both the number of edges in the graph and the number of vertices in  $R$ . The former of these is  $cl = ckr$  and the latter is  $r$ , which shows that  $OPT \leq r \max(ck, 1)$ . Therefore, by simple case analysis the approximation ratio in expectation is at least

$$\frac{1 - \exp(-ck)}{\min(ck, 1)} \geq 1 - \frac{1}{e}$$

□

However in reality, the approximation obtained by this sampling approach can be much better for certain values of  $ck$ . In particular, if  $ck > 1$ , then the approximation ratio is  $1 - \exp(-ck)$ , which approaches 1 as  $ck \rightarrow \infty$ . In particular, if  $ck = 3$ , then the solution will be at least 95% as good as the optimal solution even with our trivial bounds. Similarly, when  $ck < 1$ , the approximation ratio is  $(1 - \exp(-ck))/ck$  which also approaches 1 as  $ck \rightarrow 0$ . In particular, if  $ck = 0.1$  then the solution will be at 95% as good as the optimal solution. The case when  $ck = 1$  therefore represents the worst case outcome for this model where we only guarantee 63% optimality. The graph below shows the approximation ratio as a function of  $ck$ .

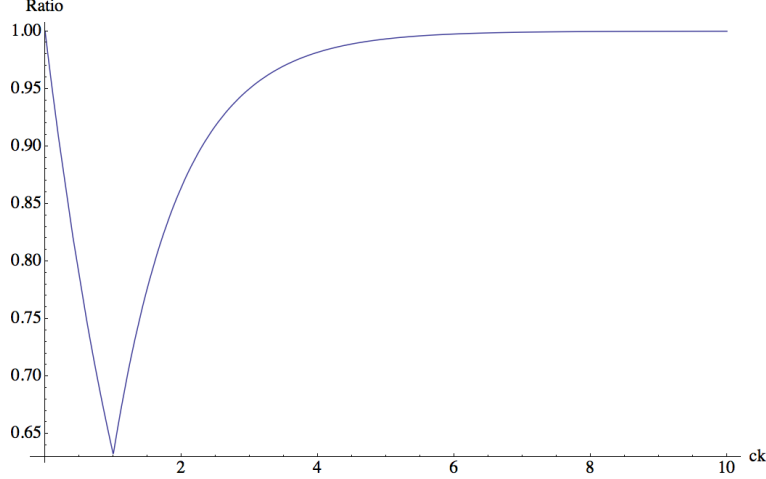


Figure 1: Approximation Ratio as a function of  $ck$

Now suppose that  $G$  is generated and  $H$  is sampled using the same processes as described above. In the next theorem, we extend the above bounds to the  $(c, a)$ -graph recommendation problem where  $a > 1$ .

**Theorem 3.** *Let  $S$  be the random variable denoting the number of vertices  $v \in R$  such that  $\deg_H(v) \geq a$ . Then*

$$\mathbb{E}[S] \geq r \left( 1 - e^{-ck + \frac{a-1}{r}} \frac{(ck)^a - 1}{ck - 1} \right)$$

where the expectation is over the randomness of  $G$  and  $H$ .

*Proof.* Let  $X_{uv}$  be the indicator variable of the event that the edge  $uv$  (note that  $u \in L$  and  $v \in R$ ) is in the subgraph that we picked and set  $X_v = \sum_{u \in U} X_{uv}$  so that  $X_v$  represents the random degree of the vertex  $v$  in our subgraph. Because our algorithm uniformly subsamples a uniformly random selection of edges, we can assume that  $H$  was generated the same way as  $G$  but sampled  $c$  instead of  $d$  edges for each vertex  $u \in L$ . So  $X_{uv}$  is a bernoulli random variable. Using the trivial bound

$\binom{n}{i} \leq n^i$  on binomial coefficients we get:

$$\begin{aligned}
\Pr[X_v < a] &= \sum_{i=0}^{a-1} \binom{cl}{i} \left(1 - \frac{1}{m}\right)^{cl-i} \left(\frac{1}{r}\right)^i \\
&\leq \sum_{i=0}^{a-1} \left(\frac{cl}{r}\right)^i \left(1 - \frac{1}{r}\right)^{cl-i} \\
&= \left(1 - \frac{1}{r}\right)^{cl-(a-1)} \sum_{i=0}^{a-1} (ck)^i \\
&\leq \left(1 - \frac{1}{r}\right)^{cl-(a-1)} \frac{(ck)^a - 1}{ck - 1} \\
&\leq e^{-ck + \frac{a-1}{r}} \frac{(ck)^a - 1}{ck - 1}
\end{aligned}$$

Letting  $Y_v = [X_v \geq a]$ , we now see that

$$E[S] = E\left[\sum_{v \in R} Y_v\right] \geq r \left(1 - e^{-ck + \frac{a-1}{r}} \frac{(ck)^a - 1}{ck - 1}\right)$$

□

Now, we can perform a similar analysis as in the previous section. In particular, if  $ck > a$ , then the problem is easy on average though we need  $ck$  to get larger than before to be close to optimal. This is in comparison to the trivial estimate of  $m$ . For a fixed  $a$ , a random solution gets better as  $ck$  increases because the decrease in  $e^{-ck}$  more than compensates for the polynomial in  $ck$  next to it. However, if  $ck < a$ , we need to use the trivial estimate of  $ckr/a$ , and the analysis from the previous section does not extend here.

In both this section and the previous one,  $ck$  is the average degree of a vertex  $v \in R$  in our chosen subgraph. Basically, what the original analysis said is that if  $ck > 1$ , then the sampling algorithm will probably cover every vertex in  $R$  since the expected degree of each vertex is large. On the other hand if  $ck$  is small ( $ck < 1$ ) then the best possible solution is obtained when none of the vertices in  $R$  has degree greater than 1. If  $ck < 1$ , then we do not cover very many vertices in  $R$ , but we also do not cover many vertices more than once. Since the optimal solution in this case was correspondingly low, our solution was good in the  $a = 1$  case. However, when  $a < 1$ , the fact that our edges are well-dispersed only hurts our solution because we need to concentrate the edges on particular  $v \in R$  that will count. The following table shows how large  $ck$  needs to be for the solution to be 95% optimal for different values of  $a$ :

| $a$  | 1    | 2    | 3    | 4     | 5     |
|------|------|------|------|-------|-------|
| $ck$ | 3.00 | 4.74 | 7.05 | 10.01 | 13.48 |

Figure 2: The required  $ck$  to obtain 95% optimality

### 3.2 Worst-Case Approximation

The results in the previous section concentrated on producing nearly optimal solutions in expectation. In this section, we will show that it is possible to obtain good solutions regardless of the

model that generated the recommendation subgraph. As usual we let  $G = (L, R, E)$  be a bipartite graph on which we would like to solve the  $(c, a)$ -graph recommendation problem and consider the following greedy algorithm. Consider each vertex in  $R$  in some arbitrary order and if there is some  $v \in R$  that has  $a$  neighbors in  $L$  all of which have degree  $< c$ , add the edges to these neighbors to  $H$ . If there are any ties about the nodes to be picked either in the selection of  $v$  or its neighbors, we can break ties arbitrarily. This algorithm has the following approximation guarantee

**Theorem 4.** *The greedy algorithm described above achieves a  $1/(a + 1)$ -approximation ratio.*

*Proof.* Let  $R_{GREEDY}, R_{OPT} \subseteq R$  be the set of vertices that have degree  $\geq a$  in the greedy and optimal solutions respectively. Note that any  $v \in R_{OPT}$  along with neighbors  $\{u_1, \dots, u_a\}$  forms a set of candidate edges that can be taken by the greedy algorithm. So we can consider  $R_{OPT}$  as a candidate pool for  $R_{GREEDY}$ . Each move that the greedy algorithm makes might make some of the candidates infeasible, but as long as the candidate pool is not depleted, the greedy algorithm can continue adding vertices to its solution. Each time the greedy algorithm claims some vertex  $v \in R$  with edges to  $\{u_1, \dots, u_a\}$ , we have obviously have to remove  $v$  from the candidate pool. If any  $u_i$  was saturated (i.e. had degree  $c$ ) in the optimal solution, we would also need to remove an arbitrary vertex  $v_i \in R$  adjacent to  $u_i$  in the optimal solution. In other words, by using an edge of  $u_i$ , we force it to not use an edge it used to some other  $v_i$ , which might cause the degree of  $v_i$  to go below  $a$ . (Note that the greedy algorithm does not actually have to be aware of the structure of optimal solution for this type of bookkeeping to go through.) Therefore, at each step of the greedy algorithm, we have to remove at most  $a + 1$  vertices from the candidate pool. Since our candidate pool has size  $OPT$ , the greedy algorithm cannot stop before it has added  $|OPT|/(a + 1)$  vertices to the solution.  $\square$

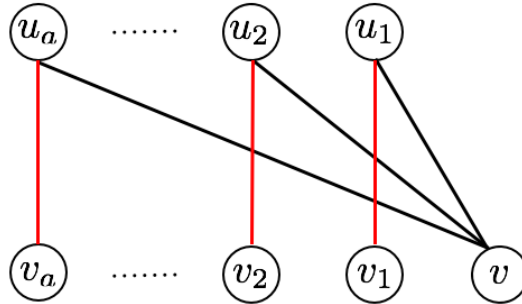


Figure 3: This diagram shows one step of the greedy algorithm. When  $v$  claims edges to  $u_1, \dots, u_a$ , it potentially removes  $v_1, \dots, v_a$  from the pool of candidates that are available. The potentially invalidated edges are shown in red.

There are two things to note about this algorithm. The first is that the solution the greedy algorithm outputs is dependent on the order the vertices in  $R$  are processed. It may be possible to improve the expected approximation ratio by permuting these vertices randomly before the algorithm runs. The second is that this approximation guarantee is as good as we can expect. In particular, if we set  $a = 1$  then we obtain the familiar  $1/2$ -approximation for matchings.

### 3.3 Comparing the Worst-Case and Average-Case Approximations

In this section, we show that the results of Section 3.1 that hold in expectation also hold in the worst case with high probability using concentration bounds. We can convert the expectation results to a lower bound that holds with high probability using concentration results like Chernoff bounds. While Chernoff bounds are usually stated for independent variables, the variant below holds for any number pairwise non-positively correlated variables.

**Theorem 5.** *Let  $X_1, \dots, X_n$  be non-positively correlated variables. If  $X = \sum_{i=1}^n X_i$ , then for any  $\delta \geq 0$*

$$\Pr[X \geq (1 + \delta)E[X]] \leq \left( \frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^{E[X]}$$

Now note that the variables  $1 - X_v$  for each  $v \in R$  are non-positively correlated. In particular, if  $N(v)$  and  $N(v')$  are disjoint, then  $1 - X_v$  and  $1 - X_{v'}$  are independent. Otherwise,  $v$  not claiming any edges can only increase the probability that  $v'$  gets an edge from any vertex  $u \in N(v) \cap N(v')$ , so the variables would be negatively correlated. Setting  $\delta = 1$  in the theorem above now proves the following:

**Theorem 6.** *The random sampling algorithm produces a solution of size at least  $r(1 - 2\exp(-ck))$  with probability at least  $1 - (e/4)^{r \exp(-ck)}$ .*

That is, the probability of producing a constant factor approximation decreases exponentially with  $r$ .

With these high probability bounds in place, it is instructive to compare the performance of the random sampling algorithm with that of the  $c = 1$ . With high probability, the approximation ratio of the randomized algorithm is  $(1 - \exp(-ck))/(1 - \exp(-dk))$  while the approximation ratio of the greedy algorithm is always at least  $1/2$ . In the following graphs  $f \geq 1$  represents a value such that  $d = fc$  and we use the values  $k = 0.1$  and  $k = 0.2$  respectively since the case that is practical for our purposes is when  $l \ll r$ . In the graphs below, the red curve shows the approximation ratio for the model used in Section 3.1 and the blue curve shows the worst case approximation ratio described in section 5. As we would expect, the sampling algorithm performs outperforms the greedy algorithm significantly when  $k$  and more importantly  $c$  gets larger.

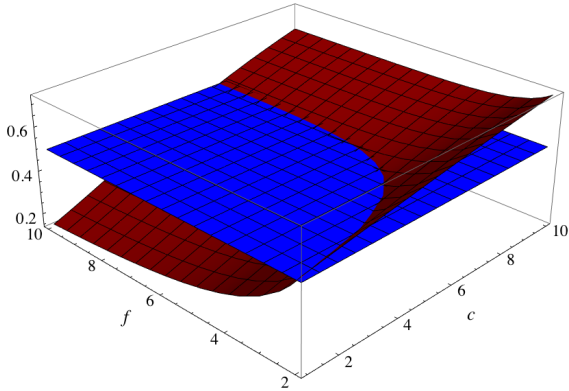


Figure 4: Approximation ratios when  $k=0.1$

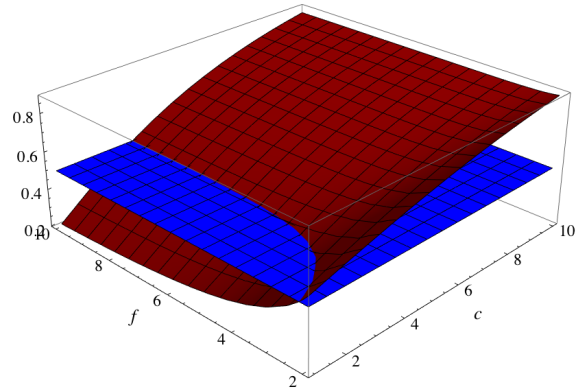


Figure 5: Approximation ratios when  $k=0.2$

### 3.4 Existence of Optimal Recommendation Subgraphs

Let  $G = (L, R, E)$  be a bipartite graph. We define a *perfect*  $(c, a)$ -recommendation on  $G$  to be a subgraph  $H$  such that  $\deg_H(u) \leq c$  for all  $u \in L$  and  $\deg_H(v) = a$  for  $\min |R|, |L|c/a$  of the  $v \in R$ . In this section we will prove a sufficient condition for perfect  $(a, a)$ -recommendation subgraphs to exist in a bipartite graph where edges are sampled uniformly and independently with probability  $p$ . Our result relies on a well-known characterization of perfect matchings [1]:

**Theorem 7.** *Let  $G = (L, R, E)$  be a bipartite graph with  $|L| = |R| = n$  and suppose that each edge is sampled uniformly and independently with probability  $p$ . If  $p \geq (\log(n) - \log \log(n)) / n$ , then  $\Pr[G \text{ has a perfect matching}] \rightarrow 1$  as  $n \rightarrow \infty$ .*

We will prove that a perfect  $(c, a)$ -recommendation subgraph exists with high probability by building it up from  $a$  matchings which must exist with high probability if  $p$  is sufficiently high. In the next theorem, we show that  $p$  only needs to be taken  $\Theta(\log(n))$  for this plan of attack to succeed. Furthermore, this decomposition into matchings motivates an algorithm that can return a solution that comes within  $(1 - \epsilon)$  of the optimal solution and can be run in almost linear time. We will state the theorem for the case when  $c < a$ , but it applies equally well to the  $a < c$  case by swapping the roles of  $L$  and  $R$  in the proof.

**Theorem 8.** *Let  $G = (L, R, E)$  be a random graph and let  $|L| = l$ ,  $|R| = r$  and  $k = l/r$ . If each edge of  $G$  is sampled with probability  $p \geq c(\log(l) - \log \log(l))/l$  and  $kc \geq a$ , then*

$$\Pr[G \text{ has a perfect } (c, a)\text{-recommendation subgraph}] \rightarrow 1$$

as  $l, r \rightarrow \infty$  together.

*Proof.* Note that given the size and the degree constraints of  $L$ , at most  $lc/a$  vertices in  $R$  can obtain degree  $a$  in a  $(c, a)$ -recommendation subgraph. We can therefore restrict  $R$  to an arbitrary subset  $R'$  of itself of size  $lc/a$ . Next, we pick an enumeration of the vertices in  $L = \{u_0, \dots, u_{l-1}\}$  put these vertices into  $a$  subsets by defining  $L_i = \{u_{(i-1)l/a}, \dots, u_{(i+c)l/a-1}\}$  for each  $1 \leq i \leq a$ . Note both  $R'$  and all of the  $L_i$  have size  $lc/a$ .

Using these new sets we defined, we define the graphs  $G_i$  on the bipartitions given by  $(L_i, R')$ . Since the sets  $L_i$  are intersecting, we cannot define the graphs  $G_i$  to be induced subgraphs. However, note that each vertex  $u \in L$  falls into exactly  $c$  of these subsets. Therefore, we can uniformly assign each edge in  $G$  to one of  $c$  graphs among  $\{G_1, \dots, G_a\}$  it can fall into, and make each of those graphs a random graph. In fact, while the different  $G_i$ s are coupled, taken in isolation we can consider any single  $G_i$  to be drawn from the distribution  $G_{lc/a, lc/a, p/c}$  since  $G$  itself was drawn from  $G_{l, r, p}$ . Since  $c/a < 1$  and  $p/c \geq (\log(l) - \log \log(l))/l$  by assumption, we can now say that

$$\Pr[G_i \text{ has no perfect matching}] = o(1)$$

Considering  $a$  to be fixed, by a union bound we can now conclude that except for a  $o(1)$  probability, each one of the  $G_i$ s has a perfect matching. By superimposing all of these perfect matchings, we can see that every vertex in  $R'$  gets degree  $a$ . Since each vertex in  $L$  is in exactly  $c$  of the  $L_i$ s, each vertex in  $L$  gets a degree  $a$ . It follows that except for a  $o(1)$  probability, there exists an  $(c, a)$ -recommendation subgraph in  $G$ .  $\square$

Using this result, we can come up with an approximation algorithm that can deliver an  $1 - \epsilon$  approximation to the  $(c, a)$ -recommendation problem almost linear time in randomly sampled dense graphs.



**Theorem 9.** *Let  $G = (L, R, E)$  be a graph where the edges are independently and uniformly with probability at least  $p \geq (\log(|L|) - \log \log(|L|))/l$ . There exists an algorithm that can with high probability output a  $(1 - \epsilon)$ -approximation in  $O(|E|a^2/\epsilon)$  time.*

*Proof.* As in the proof of the previous theorem, we can arbitrarily restrict  $R$  to a subset of size  $|L|c/a$  and divide  $L$  into sets  $L_1, \dots, L_a$  such that each vertex in  $L$  is contained in exactly  $c$  of the sets. Using the previous theorem, we know that each of the graphs  $G_i$  has a perfect matching with high probability. These perfect matchings can be approximated to a  $1 - \epsilon/a$  factor by finding matchings that don't have augmenting paths of length  $\geq 2a/\epsilon$ . This can be done for each  $G_i$  in  $O(|E|a/\epsilon)$  time. Furthermore, the union of unmatched vertices makes up an at most  $a \cdot (\epsilon/a)$  fraction of  $R'$ . This proves the result.  $\square$

While this result is a heavier weight solution, rephrasing the problem in terms of matchings enabled us to take advantage of the well-understood combinatorial properties of matchings

## 4 More Expressive Models of Recommendation Graphs

Having described the fixed-degree model in detail above, we can now use it as a template for the analysis of three more richer and more expressive models. In Section 4.3, we generalize the fixed degree model to a model where the graph is fixed and the edges can weights. In Section 4.1, we analyze the *hierarchical tree model* which aims to model recommendation trees that arise from relationships in a product hierarchy. Finally, in Section 4.2 we study the *cartesian product model* which aims to model recommendation graphs which arise from combining the relevancy advice of several different algorithms.

### 4.1 Hierarchical Tree Model

In this section we explore the hierarchical tree model. We will assume that we are given a bipartite graph  $G = (L, R, E)$ . The vertex sets  $L$  and  $R$  are the leaf sets of two full binary trees  $T_L$  and  $T_R$  of depth  $D$  where there is a one-to-one correspondence between the subtrees of these two trees. We also assume that each branching in both  $T_L$  and  $T_R$  splits the nodes evenly into the two subtrees. As in the previous sections, we set  $|L|/|R| = k$ , but we also note that this ratio still holds if we take any subtree on the left and its corresponding subtree on the right. By abuse of notation, we will use a subtree and its leaf set interchangeably. The trees are fixed in advance, but  $G$  is generated probabilistically according to the following procedure. Let  $u \in L$  and  $T_L^0, \dots, T_L^{D-1}$  be the subtrees it belongs at depths  $0, \dots, D-1$ . Also, let  $T_R^0, \dots, T_R^{D-1}$  be the subtrees on the right that correspond to these trees on the left. We let  $u$  make an edge to  $d_{D-1}$  of the vertices in  $T_R^{D-1}$ ,  $d_{D-2}$  edges to the vertices in  $T_R^{D-1} \setminus T_R^{D-2}$  and so on. Each vertex is chosen with uniform probability and we let  $d = d_0 + \dots + d_{D-1}$ .

Our goal now is to find a  $b$ -matching in this graph that is close to optimal in expectation. That is, our degree upper bound on vertices in  $L$  is  $c$  and the degree lower bound on vertices in  $L$  is 1. Let  $c = c_0 + \dots + c_{D-1}$  similarly to how we defined  $d$ . To combine the analysis of the randomness of the algorithm and the randomness of the graph, the algorithm will pick  $c_i$  edges uniformly from among the  $d_i$  edges going to each level of the subtree. This enables us to think of the subgraph our algorithm outputs as being generated by the graph generation process, but with fewer neighbors selected for each as in the previous sections. With this model and parameters in place, we can prove the following theorem.

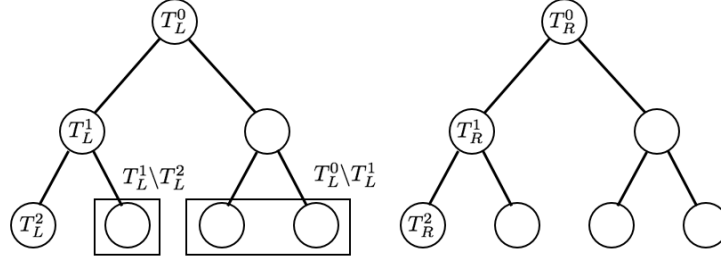


Figure 6: This diagram shows the notation we use for this model and the 1-to-1 correspondance of subtrees

**Theorem 10.** *Let  $S$  be the subset of edges  $v \in R$  such that  $\deg_H(v) \geq 1$ . Then*

$$E[S] \geq r(1 - \exp(-ck))$$

where the expectation is taken over  $G$  and  $H$ .

*Proof.* Let  $v \in R$  and let  $T_L^{D-1}, T_L^{D-2} \setminus T_L^{D-1}, \dots, T_L^0 \setminus T_L^1$  be the sets it can take edges from. Since  $T_L$  and  $T_R$  split perfectly evenly at each node the vertices in these sets will be choosing from  $r_{D-1}, r_{D-1}, r_{D-2}, \dots, r_1$  vertices in  $R$  for neighbors respectively, where  $r_i$  is the size of subtree of the right tree rooted at depth  $i$ . Furthermore, each of these sets described above have size  $l_{D-1}, l_{D-1}, l_{D-2}, \dots, l_1$  respectively, where  $l_i$  is the size of a subtree of  $T_L$  rooted at depth  $i$ . It follows that the probability that  $v$  does not receive any edges at all is at most

$$\begin{aligned} \Pr[\neg X_v] &= \left(1 - \frac{1}{r_{D-1}}\right)^{c_0 l_{D-1}} \prod_{i=1}^{D-1} \left(1 - \frac{1}{r_i}\right)^{c_{D-i} l_i} \\ &\leq \exp\left(-\frac{l_{D-1}}{r_{D-1}} c_0\right) \prod_{i=1}^{D-1} \exp\left(-\frac{l_i}{r_i} c_{D-i}\right) \\ &= \exp(-(c_0 + \dots + c_{D-1})k) \\ &= \exp(-ck) \end{aligned}$$

Since this is an indicator variable, it follows that

$$E[S] = E\left[\sum_{v \in R} X_v\right] \geq r(1 - \exp(-ck))$$

□

Note that this is the same result as we obtained for the fixed degree model in Section 3.1. In fact, the approximation guarantees when  $ck \ll 1$  or  $ck \gg 1$  hold exactly as before.

The sampling of  $H$  can be done algorithmically because we separated out the edge generation process at a given depth from the edge generation process at deeper subtrees. There is no ambiguity as to why an edge is in the underlying graph. That is, if we superimpose  $T_L$  and  $T_R$ , then an edge between  $u_l \in L$  and  $v_r \in R$  must have come from an edge generated at the lowest common ancestor of  $u_l$  and  $v_r$ . So the algorithm can actually sample intelligently and in the same way that the graph was generated in the first place. Also note that we do not have to assume that the trees  $T_L$  and  $T_R$  are binary. We only need the trees to be regular and evenly divided at each vertex since the proof only relies on the proportions of the sizes of the subtrees in  $T_L$  and  $T_R$ .

## 4.2 Cartesian Product Model

We can extend the analysis in Section 3.1 in a way orthogonal to the hierarchical tree model as follows. We assume that  $L$  has been partitioned into  $t$  subsets  $L_1, \dots, L_t$  and that  $R$  has been partitioned into  $t'$  subsets  $R_1, \dots, R_{t'}$ . For convenience, we let  $|L_i| = l_i$  and  $|R_i| = r_i$ . Given this suppose that for each  $1 \leq i \leq t$  and each  $1 \leq j \leq t'$ ,  $G[L_i, R_j]$  is an instance of the Fixed Degree Model with  $d = d_{ij}$ . We assume that for all  $i$ , we have  $\sum_{j=1}^{t'} d_{ij} = d$  for some fixed  $d$ . Also assume that we have fixed in advance  $c_{ij}$  for each  $1 \leq i \leq t$  and  $1 \leq j \leq t'$  that satisfy  $\sum_{j=1}^{t'} c_{ij} = c$  for all  $i$  for some fixed  $c$ . To sample  $H$  from  $G$ , we sample  $c_{ij}$  neighbors for each  $u_i \in L_i$  from  $R_i$ . Letting  $S$  be the set of vertices in  $v \in R$  that satisfy  $\deg_H(v) \geq 1$ , we can show the following:

**Theorem 11.** *With  $S$ ,  $G$  and  $H$  defined as above, we have*

$$E[S] \geq r - \sum_{j=1}^{t'} r_j \exp \left( - \sum_{i=1}^t c_{ij} \frac{l_i}{r_j} \right)$$

where the expectation is over  $G$  and  $H$ .

*Proof.* Let  $v_j \in R_j$  be an arbitrary vertex and let  $X_{v_j}$  be the indicator variable for the event that  $\deg_H(v_j) \geq 1$ . The probability that none of the neighbors of some  $u_i \in R_i$  is  $v_j$  is exactly  $(1 - \frac{1}{r_j})^{c_{ij}}$ . It follows that the probability that the degree of  $v_j$  in the subgraph  $H[L_i, R_j]$  is 0 is at most  $(1 - \frac{1}{r_j})^{c_{ij} l_i}$ . Considering this probability over all  $R_j$  gives us:

$$\Pr[X_{v_i} = 0] = \prod_{i=1}^t \left( 1 - \frac{1}{r_j} \right)^{c_{ij} l_i} \leq \exp \left( - \sum_{i=1}^t c_{ij} \frac{l_i}{r_j} \right)$$

By linearity of expectation  $E[S] = \sum_{i=1}^{t'} r_i E[X_{v_i}]$ , so it follows that

$$E[S] \geq \sum_{j=1}^{t'} r_j \left( 1 - \exp \left( - \sum_{i=1}^t c_{ij} \frac{l_i}{r_j} \right) \right) = r - \sum_{j=1}^{t'} r_j \exp \left( - \sum_{i=1}^t c_{ij} \frac{l_i}{r_j} \right)$$

□

This model is interesting because it can capture a broader set of recommendation subgraphs than the fixed degree model. However, it is difficult to estimate how good a solution will be without knowing the sizes of the sets in the partitions. However, we can note that we obtain the approximation guarantee of  $(1 - \exp(-ck))$  provided that  $l_i/r_j = k$  for all  $i$  and  $j$  where  $k$  is some fixed constant. Another interesting point about this model and the algorithm we described for sampling  $H$  is that we are free to set the  $c_{ij}$  as we see fit. In particular,  $c_{ij}$  can be chosen to maximize the approximation guarantee in expectation we obtained above using gradient descent or some other first order method prior to running the recommendation algorithm to increase the quality of the solution.

## 4.3 Weighted Model

The fixed degree model of Section 3.1 is a simple and convenient model, but the assumption that all recommendations hold the same weight is unrealistic. This motivates fixing the graph to be the complete bipartite graph  $K_{l,r}$ , and giving the edges i.i.d weights with mean  $\mu$ . We modify the

objective function accordingly, so that we count only the vertices in  $R$  which have weight  $\geq 1$ . If we assume that  $ck\mu \geq 1 + \epsilon$  for some  $\epsilon > 0$ , then naive the solution sampling solution we outlined in Section 3.1 still performs exceptionally well. Letting  $S$  be the size of the solution produced by this algorithm we have:

**Theorem 12.** *Let  $G = K_{l,r}$  be a bipartite graph where the edges have i.i.d. weights and come from a distribution with mean  $\mu$  that is supported on  $[0, b]$ . If the algorithm from Section 3.1 is used to sample a subgraph  $H$  from  $G$ , then*

$$E[S] = \sum_{v \in R} E[X_v] = r \left( 1 - \exp \left( -\frac{2l\epsilon^2}{b^2} \right) \right)$$

*Proof.* For each edge  $uv \in G$ , let  $W_{uv}$  be its random weight,  $Y_{uv}$  be the indicator for the event  $uv \in H$  and define  $X_{uv} = Y_{uv}W_{uv}$ . Since weights and edges are sampled by independent processes, we have  $E[X_{uv}] = E[W_{uv}]E[Y_{uv}]$  for all edges. Since  $c$  edges out of  $r$  are picked for each vertex,  $E[Y_{uv}] = \frac{c}{r}$ , so  $E[X_{uv}] = \frac{c}{r}\mu$ . Therefore, the expected weight coming into a vertex  $v \in R$  would be

$$E[X_v] = \sum_{u \in L} E[X_{uv}] = \frac{cl\mu}{r} = ck\mu$$

However,  $X_{uv}$  for each  $u$  are i.i.d random variables. Since by assumption  $ck\mu = 1 + \epsilon$ , by a Hoeffding bound we can obtain:

$$\Pr[X_v \leq 1] = \Pr[X_v - E[X_v] \geq \epsilon] \leq \exp \left( -\frac{2l\epsilon^2}{b^2} \right)$$

By linearity of expectation we can now get the result in the theorem

$$E[S] = \sum_{v \in R} E[X_v] = r \left( 1 - \exp \left( -\frac{2l\epsilon^2}{b^2} \right) \right)$$

□

There are two things to note about this variant. The first is that since the variables  $X_v$  are negatively correlated, our results in 3.2 can be readily extended to the results of this section. The second is that the condition that  $W_{uv}$  are i.i.d is not necessary to obtain the full effect of the analysis. Indeed, the only place in the proof where the fact that  $W_{uv}$  are i.i.d is when we argued that  $X_{uv}$  is large with high probability by a Hoeffding bound. For the bound to apply, it's sufficient to assume that  $W_{uv}$  for all  $v$  are independent. In particular, it's possible that  $W_{uv}$  for all  $u$  are inter-dependent. This allows us to assume an weight distribution that depends on the strength of the recommender and the relevance of the recommendation separately.

## 5 Future Directions

In this paper we proposed several different models for explaining how recommendation subgraphs arise probabilistically and how optimization problems on these graphs can be solved, but there is much more that can be done.

1. In practice, there might exist several different recommendation subgraphs based on different features. For example, two people might be related because they went to the same school,

or because they live in the same city, or because they would complete a large number of triangles, etc. It's worthwhile to investigate how such graphs can be combined into one, or how an optimization problem can be solved using all such recommendation subgraphs simultaneously.

2. We should devise metrics that can evaluate how well a recommendation subgraph fits a given model and conduct some parameter fitting experiments to see how well actual recommendation subgraphs which arise in practice fit our models.
3. We should implement the sampling and the greedy algorithms given in the paper to see if they can solve to near optimality the graph recommendation problems that arise in practice.

## References

- [1] Svante Janson, Tomasz Luczak, and Andrzej Rucinski. *Random Graphs*. Wiley Series in Discrete Mathematics and Optimization. Wiley, 2011.