Social Computing **CS60017** **Autumn Semester, 2014**
End-Sem **Maximum Marks: 90** **Time Limit: 3 Hours**

This exam contains 4 pages (including this cover page) and 8 problems.

You may *not* use your books or notes in this exam.

All the sub-parts for a given question *must be answered at one place only.*

1. (10 points) Answer the following:
   (a) (3 points) In the context of the Category Game, explain with examples of a few game steps how a word contagion, i.e., expansion of the linguistic category boundaries can take place.
   (b) (3 points) You are given a set of m objects that is divided into K groups, where the i^{th} group is of size m_i. If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement.) – (i) We randomly select n^*m_i/m elements from each group OR (ii) We randomly select n elements from the data set, without regard for the group to which an object belongs.
   (c) (4 points) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming distance or Jaccard similairty, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance or the Jaccard coefficient? Explain. (Note that two human beings share $> 99.9\%$ of the same genes.)

2. (17 points) Answer the following:
   (a) (4 points) Given two documents D1 = {"His brow trembled"} and D2 = {"The brown emblem"}, compute all k-shingles for each document with $k = 4$. Compute the Jaccard Similarity of the two sets of k-shingles for D1 and D2.
   (b) (4 points) Show that for any two sets S_i and S_j, the probability that $h(S_i) = h(S_j)$ is equal to the Jaccard similarity of S_i and S_j where $h()$ is the standard minhash.
   (c) (4 points) Construct the minhash signature matrix created from the table given below using $n = 2$ permutations (using functions $h_1 = (x + 1) \bmod 5$ and $h_2 = (3x + 1) \bmod 5$).

Element	S_1	S_2	S_3	S_4
a	1	0	1	1
b	1	0	1	0
c	0	1	0	1
d	1	0	1	0
e	0	0	1	0

   (d) (5 points) Let s be the Jaccard similarity of two documents. Compute the probability that the minhash signature matrix agrees in all the rows of at least one band for these two documents where b is the number of bands and r is the number of consecutive rows in each band.

3. (8 points) Consider the two tables given below. The first table shows the number of common predictions made by various link prediction methods for a co-authorship dataset of physics authors. The second table shows the number of common correct predictions made by these set of methods. From the first table which are the closest methods? From the second table which methods seem to be close? Which among these two results do you consider more significant? Which method appears to be the (i) best and (ii) worst? Explain.

	Adamic/Adar	Katz clustering	common neighbors	hitting time	Jaccard's coefficient	weighted Katz	low-rank inner product	rooted Pagerank	SimRank	unseen bigrams
Adamic/Adar	1150	638	520	193	442	1011	905	528	372	486
Katz clustering		1150	411	182	285	630	623	347	245	389
common neighbors			1150	135	506	494	467	305	332	489
hitting time				1150	87	191	192	247	130	156
Jaccard's coefficient					1150	414	382	504	845	458
weighted Katz						1150	1013	488	344	474
low-rank inner product							1150	453	320	448
rooted Pagerank								1150	678	461
SimRank									1150	423
unseen bigrams										1150

	Adamic/Adar	Katz clustering	common neighbors	hitting time	Jaccard's coefficient	weighted Katz	low-rank inner product	rooted Pagerank	SimRank	unseen bigrams
Adamic/Adar	92	65	53	22	43	87	72	44	36	49
Katz clustering		78	41	20	29	66	60	31	22	37
common neighbors			69	13	43	52	43	27	26	40
hitting time				40	8	22	19	17	9	15
Jaccard's coefficient					71	41	32	39	51	43
weighted Katz						92	75	44	32	51
low-rank inner product							79	39	26	46
rooted Pagerank								69	48	39
SimRank									66	34
unseen bigrams										68

4. (5 points) What are the 5 components of an account hierarchy in keyword based computational advertising? Which component stores the following information related to budget and targeting? Which component is shown to the user? In this component, which component is not explicitly visible on the web page?

5. (15 points) Mybook is a new social networking platform. They have 10 Million active users and they are interested in monetizing on their platform. One of the key areas where individuals have posts are related to cars. There are 2 advertisers who are interested in reaching out to

the users on MyBook. MyBook has only one slot on their page where they can show ads. They are auctioning this slot on their car related pages.

Once a user clicks on an ad, we assume that the click leads to an eventual purchase. The two advertisers value the clicks to be v_1 and v_2 respectively. MyBook estimates the probability of the click from each ad to be p_1 and p_2 respectively. Each of the advertiser place their bids (b_1 and b_2) for winning the ad slot.

In order to rank the ads, MyBook uses the product of bid and the probability of a click as a scoring function (F). The ads are ranked based on this scoring function (F_1 and F_2 for each advertiser respectively) and the one with the higher score wins the auction. For this problem, the cost paid by the advertiser is Cost per click = (Score_low/Score_High) * bid of the advertiser with higher score.

Under GSP with truthful bidding, when $v_1 > v_2$ and under the assumption that $p_1 > p_2$:

(a) (4 points) In a particular auction, $p_1 = 0.1$, $p_2 = 0.01$; $b_1 = \$100$; $b_2 = \$10$; who will win the auction and what price does the winner have to pay?

(b) (4 points) What is the relationship between the two advertisers with respect to their scoring function (F).

(c) (5 points) Show how the cost per click under this auction would be always less than or equal to the value of the click associated by the advertiser.

(d) (2 points) What would be the bid value (safe bidding strategy that will ensure win yet limit their cost) for the advertiser who wins under equilibrium (i.e., after participating in the auction say 1000 times).

6. (12 points) Answer the following:

(a) (6 points) Suppose you are using Gibbs sampling to estimate the distributions, θ and β for topic models. The underlying corpus has 5 documents and 5 words, {*River*, *Stream*, *Bank*, *Money*, *Loan*} and the number of topics is 2. At certain point, the structure of the documents looks like the following Table. For instance, the first row indicates that the document 1 contains 4 instances of word 'Bank', 6 instances of word 'Money' and 6 instances of word 'Loan'. Black and white circles denote whether the word is currently assigned to topics t_1 and t_2 respectively.
Use this structure to estimate ${\beta_{MONEY}}^{(2)}$ and ${\beta_{BANK}}^{(1)}$ at this point. You can take the values of η and α to be 0.1 each.

Doc. Id	River	Stream	Bank	Money	Loan
1			● ● ●●	● ● ● ● ●●	● ● ● ● ●●
2			● ● ● ○ ●	● ● ● ● ● ● ●	● ● ●●
3	○	○ ○ ○	● ○ ○ ○ ●○	● ● ●●	● ● ●
4	○ ○ ○ ○ ○○	○ ○ ○	● ○ ○ ○ ○○		
5	○○	○ ○ ○ ○ ○○	○ ○ ○ ○ ○○		

(b) (6 points) Explain briefly how the relational topic model (RTM) modifies the basic settings of LDA to be able to use it for the task of link prediction. Is there any advantage of using RTM than directly predicting using the topic proportions obtained from the basic LDA model?

7. (15 points) Consider the following ratings provided by 5 users, Alice, User1 - User4, to 5 items, Item1 to Item5.

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

(a) (5 points) Use item-based collaborative filtering to predict the rating that Alice will give for Item5.

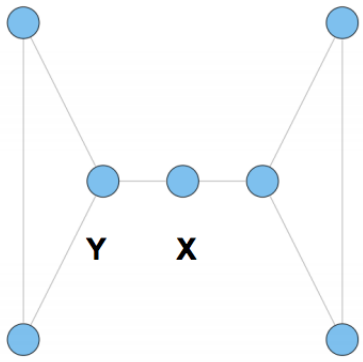
(b) (6 points) Assume that there is an underlying social network between these 5 users, which is given by the following adjacency list. The network is directed.

Alice, User1 Alice, User2 Alice, User3
User1, User3 User1, User4
User2, User3 User2, User1
User3, User4 User3, User2
User4, User 3

Also, assume that the ratings given by the users to various items are same as in the above matrix, *except that we do not have the ratings provided by User1 and User2 to Item5 anymore*. Suppose you are using the TrustWalker method to predict the rating of Item5 by the user 'Alice'. Assuming that at each step, you can choose any of the direct neighbors with equal probability, find out the probability that the random walk will continue for more than 1 step.

(c) (4 points) How would you incorporate this social network information in the matrix factorization model? Assume that the trust values between all connected user pairs are identical.

8. (8 points) Consider the following network.



Among nodes X and Y, which node has

(a) higher betweenness centrality

(b) higher closeness centrality

Report the values.