Social Computing	**CS60017**	**Autumn Semester, 2015**
End-Sem	**Maximum Marks: 80**	**Time Limit: 3 Hours**

This exam contains 4 pages (including this cover page) and 10 problems.

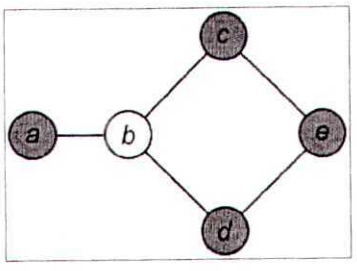
You may *not* use your books or notes for this exam. Be *precise* in your answers. All the *sub-parts* of a problem should be answered at *one place* only. On multiple attempts, *cross* any attempt that you do not want to be graded for.

There are no clarifications. In case of doubt, you can take a valid assumption, state that properly and continue.

1. (8 points) Answer the following questions:

   (a) Contrast regular and random graphs using clustering and path length. How does it explain the clustering and path length, observed in social networks? (3 points)

   (b) Find the centralization of the following network using betweenness centrality. (5 points)



2. (7 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.

   Suppose that you are at Document 3, at the first occurrence of word Bank (indicated by black circle in the table). Compute the probability distribution from which you will sample a topic for this word. 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	○○	○○○○○○	○○○○○○		

3. (6 points) Give the generative model of LDA and explain how do you modify the basic settings of this model to

i). Incorporate Scholarly Impact ii). Give a supervised topic model

What would be an alternative to Supervised LDA and why the supervised model is preferable?

4. (14 points) Consider the following table that describes the ratings given by 4 different users to 4 items. Question mark denotes the absence of rating.

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

(a) Use user-based collaborative filtering to predict the rating of User1 for Item4. (4 points)

(b) Suppose you perform SVD over the user-item interaction matrix and the matrices U_k and V_k, obtained are as follows:

U_k	Dim1	Dim2	V_k	Dim1	Dim2
User1	0.31	0.93	Item1	-0.44	0.58
User2	0.70	-0.06	Item2	-0.57	-0.66
User3	-0.44	0.23	Item3	0.06	0.26
User4	0.47	-0.30	Item4	0.38	0.18

Also, assume that the diagonal elements of the matrix Σ, in no particular order, are $\{1.5, 0.5, 4.5, 2\}$. Predict the rating that User1 will give to Item4. (5 points)

(c) Assume that there is an underlying social network between these 4 users, which is given by the following adjacency list. The network is directed.

User1, User3 User2, User1
User2, User3 User2, User4
User3, User2 User4, User1

Suppose that the interest for the four users are given as follows:

User1 (Books, Music, Software)
User2 (Books, Music)
User3 (Music, Software)
User4 (Books, Software)

How would you formulate the objective function for circle-based Social recommendation? Mention it for category 'Books' with the only variable being the latent factor for the users and items. (5 points)

5. (15 points) Hashtags are an integral part of social media content. They are used for explicitly marking the topic or an opinion/sentiment of a post. They facilitate searching through similar posts on a topic. However, users are not very consistent in annotating their posts with hashtags. Sometimes no hashtags are provided, while sometimes only a subset of the meaningful or applicable hashtags are provided. Therefore, automatic prediction of hashtags is a useful application. Let us assume that there is a tweet t coming from a user u. t may or may not contain any hashtag. The task is to predict all the hashtags that are applicable to t.

(a) "Predicting ALL the applicable hashtags of t is neither possible nor a useful proposition." Do you agree? If you do, then explain why and then come up with a more realistic

proposition (i.e., problem definition). If you don't, then show how one can practically predict and use all possible hashtags applicable to t. (2 points)

(b) Suppose you want to use a machine learning technique to solve the hashtag prediction problem. How would you go about collecting training and test data with minimal effort? (2 points)

(c) What features would you use for solving this problem? Think in terms of the features related to the content of t, the user u, and the relationship between t and all other tweets (and maybe u with all other users and tweets) (4 points)

(d) Suppose in your test data, t is associated with hashtags h_1, h_2 and h_3. And a machine learning algorithm suggested the tags g_1, g_2, g_3, g_4 and g_5 for t. Assume that all the h_i's and g_i's are unique, but h_i can be similar to g_j (where i may or may not be equal to j). How would you define the goodness of your system? Propose a metric based on h_i's and g_i's such that a better machine learning system would be expected to have a higher value for the metric. (3 points)

(e) However, for an ideal evaluation of the system one cannot rely on a static test data. Explain why? And propose a better method for evaluation. (4 points)

6. (9 points) Consider the following dataset and answer the questions.

Data Set	#	SSN	Name	DOB	Sex	ZIP
Set A	1	000956723	Smith, William	1973/01/02	Male	94701
	2	000956723	Smith, William	1973/01/02	Male	94703
	3	000005555	Jones, Robert	1942/08/14	Male	94701
	4	123001234	Sue, Mary	1972/11/19	Female	94109
Set B	1	000005555	Jones, Bob	1942/08/14		
	2		Smith, Bill	1973/01/02	Male	94701

(a) What are the potentially similar entities in this dataset? (2 points)

(b) Outline a possible key that could be used to execute the sorted-neighborhood method. (3 points)

(c) Explicitly list the keys and show how this can identify duplicates through a run of the sorted-neighborhood. (4 points)

7. (5 points) Compute the (i) Jaro and (ii) Jaro-Winkler similarity for the strings s_1 = DIXON and s_2=DICKSONX. Assume $p = 0.1$ for the computations.

8. (4 points) Consider the sets $S_1 = \{1, 2, 5\}$, $S_2 = \{3\}$, $S_3 = \{2, 3, 4, 6\}$ and $S_4 = \{1, 4, 6\}$ and $U = \{1, 2, 3, 4, 5, 6\}$.

   (a) Construct the characteristic matrix from this set. (2 points)

   (b) Suppose after permutation the row numbers change as follows $2 \rightarrow 1$, $5 \rightarrow 2$, $6 \rightarrow 3$, $1 \rightarrow 4$, $4 \rightarrow 5$, $3 \rightarrow 6$. Compute the minhash for each set. (2 points)

9. (2 points) Describe briefly why is the problem of link prediction thought to be hard.

10. (10 points) The function $p = 1 - (1 - s^r)^b$ gives the probability p that two minhash signatures that come from sets with Jaccard similarity s will hash to the same bucket at least once, if we use an LSH scheme with b bands of r rows each. For a given similarity threshold s, we want to choose b and r so that $p = 1/2$ at s. Suppose that signatures have length 24, which means we can pick any integers b and r whose product is 24. That is, the choices for r are 1, 2, 3, 4, 6, 8, 12, or 24, and b must then be $24/r$.

    (a) If $s = 1/2$, determine the value of p for each choice of b and r. Which r would you choose, if $1/2$ were the similarity threshold? (5 points)

    (b) For each choice of b and r, determine the value of s that makes $p = 1/2$ (5 points)