**EE 219**

**Graphs and Network Flows**

Pinggyuan Yue

Homework 1 Report

Pingyuan Yue: 504737715

Ke Xu: 604761427

Yuanyi Ding: 404773978

2/12/2017

Table of Contents

[**1. Introduction** 3](#_Toc474858186)

[**2. Dataset and Problem Statement** 3](#_Toc474858187)

[(a) Loading data and histogram plotting 3](#_Toc474858188)

[**3. Modeling Text Data and Feature Extraction** 4](#_Toc474858189)

[(b) Creating Bags of Words and TF-IDF Representation 4](#_Toc474858190)

[(c) 10 Most Significant Features Based on TF-ICF 5](#_Toc474858191)

[**4. Feature Selection** 7](#_Toc474858192)

[(d) LSI Transform 7](#_Toc474858193)

[**5. Learning Algorithms** 8](#_Toc474858194)

[(e) Hard Margin SVM classification 8](#_Toc474858195)

[(f) Soft Margin SVM Classification 10](#_Toc474858196)

[(g) Naive Bayes Algorithm 11](#_Toc474858197)

[(h) Logistic Regression Classification 13](#_Toc474858198)

[(i) Logistic Regression with L-1 and L-2 Norm 15](#_Toc474858199)

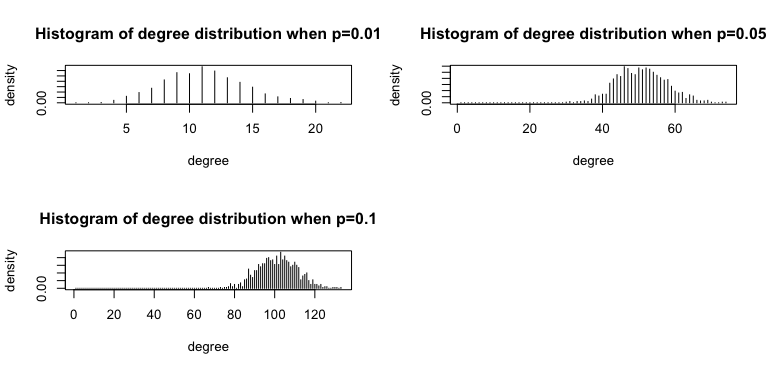
[**6. Multiclass Classification** 17](#_Toc474858200)

[(j) Multiclass Classification with Naive Bayes Algorithm and SVM Classification 17](#_Toc474858201)

# Create Random Networks

**Part (a):**

We are asked to create three undirected random networks with 1000 nodes, and the probability p for connecting two arbitrary vertices 0.01, 0.05 and 0.1 respectively. Below are three figures of degree distribution:



Figrue1.1: degree distribution for different value of p

We observed from the graphs above that the networks trend to have high degree if the probability p for connecting two arbitrary vertices is high.

**Part (b):**

Due to fact that our graphs are randomly generated, we decide to repeat the process for 100 times and compute the average for connectivity and diameter. The results are shown as following:

|  |  |  |  |
| --- | --- | --- | --- |
|  | P = 0.01 | P = 0.05 | P = 0.1 |
| Connectivity | 0.99 | 1 | 1 |
| Diameters | 5.34 | 3 | 3 |

Table 1.1: connectivity and diameters for different value of p

**Part (c):**

We are asked to find out a value of , so that any networks we generated with a p that is smaller than are disconnected, and any networks generated with a p that is larger than are connected. Considering the results in part b, we decide to perform a binary search between 0 and 0.01 to find the . The algorithm stops whenever the graph we generate is disconnected with possibility of (p - ε) and is connected with possibility of (p + ε). That value of p is the we try to find. The we find is 0.00728.

**Part (d):**

We are asked to derive the value of analytically. We assume the with certain value of p, we expect the find one single isolated node. The expectation of finding this node can be represented as:

1 = , where N = 1000

p =

The p we calculate from the above equation is equal to 0.00688, which is close to what we get from part c.