Assignment 2: Link Analysis & Prediction; Community Detection

Ahmad Baig
York University
Mississauga, Ontario, Canada
ahmad85@my.yorku.ca

ABSTRACT

In Assignment 2 I used the DBLP co-authorship data-set to generate three different kinds of graphs. I was particularly interested in co-authorship that took place in the years 2005 and 2006. After I extracted the data and made graphs out of them I used PageRank to obtain the 50 most important authors of that particular year. Then I found out the 20 most important edges by measuring edge-betweeness scores.I tested out several prediction algorithms to confirm the precision of the various algorithms. I tested out the prediction algorithms by obtaining the edges that were friends of friends(FoF's) I then ran the algorithms on the FoF's and made sure that the top K results based on predicted links where in the the graph for the year 2006.

WARNING

IN THIS ASSIGNMENT I HAD COMPUTATION ISSUES DUE TO THE SLOWNESS OF MY COMPUTER FOR EDGE BETWEENESS IN PART B ii. I SET K=5 IN ORDER TO COMPUTE IT IN A REASONABLE TIME ON MY COMPUTER. ALSO FOR THE GAVIN NEW MIN SECTION PART D I SET THE DEGREE GREATER THAN 30 SO THAT I COULD COMPUTE IT ON MY COMPUTER. MY APOLOGIES.

PART A. TEMPORAL GRAPHS

In this section I have extracted authorship events that took place in the years 2005 and 2006. I did this to lessen the amount of time it took to look through the data and generate the graphs that were needed.

dblp2005: undirected unweighted graph that represents the DBLP co-authorships of the year 2005

Graph generation process and results for dblp2005. First I created dblp2005 which was an undirected unweighted graph that represented the DBLP co-authorships of the year 2005. After the giant connected component was obtained you can see that these where the number of nodes and edges:

Number of nodes: 106943

Number of edges: 300043

dblp2006: undirected unweighted graph that represents the DBLP co-authorships of the year 2006

Graph generation process and results for dblp2006. Secondly, I created dblp2006 which was an undirected unweighted graph that represented the DBLP co-authorships of the year 2006. After the giant connected component was obtained you can see that these where the number of nodes and edges:

Number of nodes: 123808

Number of edges: 356968

dblp2005w: undirected weighted graph that represents the DBLP co-authorships of the year 2005

Graph generation process and results for dblp2005w. Thirdly, I created dblp2005w which was an undirected weighted graph that represented the DBLP co-authorships of the year 2005. I did this by setting every edge to weight of one if that edge already existed in the graph as I inserted it into the the graph I would increment the edge weight by one.

After the giant connected component was obtained you can see that these where the number of nodes and edges:

Number of nodes: 106943

Number of edges: 300043

PART B. NODE AND EDGE IMPORTANCE IN GRAPHS

In this section I created I found the PageRank and edge betweeness scores for all of the nodes and edges respectively, I then sorted them by greatest to least. Finally I selected top 50 PageRank scores and top 20 edge betweeness scores.

dblp2005 PageRank and Edge Betweeness scores

TOP 50: PageRank scores for dblp2005 graph.

('Wen Gao', 0.0001419310322458536)
('Xin Li', 0.0001330668725233402)
('Wei Zhang', 0.0001183709570000217)
('Chin-Chen Chang', 0.00011228187096190702)
('Wei Wang', 0.00010495905373283141)
('Guanrong Chen', 0.00010241545494191464)
('Zhaohui Wu', 9.972482112399877e-05)
('Wei-Ying Ma', 9.740727211556884e-05)
('Bin Wang', 9.726301353073489e-05)
('Samuel Pierre', 9.705863952617082e-05)
('Hans-Peter Seidel', 9.630692366059367e-05)
('Francky Catthoor', 9.5096877551315 48e-05)
('Yong Zhang', 9.502779502456232e-05)
('Fang Liu', 9.291312733184175e-05)
('Witold Pedrycz', 9.159654224986458e-05)
('Wei Liu', 9.088044455802498e-05)
('Alberto L. Sangiovanni-Vincentelli', 9.062998434173117e-05)
('Licheng Jiao', 9.054039000281964e-05)
('Yan Zhang', 8.964011202963375e-05)
('Xiang Li', 8.924949886585826e-05)
('Mario Piattini', 8.77030168704954e-05)
('Yvon Savaria', 8.759316637739888e-05)
('Yan Li', 8.757045759012319e-05)
('David J. Evans', 8.646019165514895e-05)
('Wei Zhao', 8.631862736191031e-05)
('Lei Li', 8.521097914198228e-05)
('Minglu Li', 8.385477457657978e-05)
('Jing Li', 8.364797920281 22e-05)
('Qing Li', 8.346925817780209e-05)
('Xianlong Hong', 8.280253482108303e-05)
('Jun Li', 8.255650675543822e-05)
('Mario Gerla', 8.151790992165258e-05)
('Yu Chen', 8.098964142814071e-05)
('Yang Xu', 8.094342574702077e-05)
('Tharam S. Dillon', 8.027406115457522e-05)
('Jiannong Cao', 7.996882944317616e-05)
('Elisa Bertino', 7.975921070453302e-05)
('Yan Wang', 7.953052894162 08e-05)
('Philip S. Yu', 7.904077202276816e-05)
('Azzedine Boukerche', 7.863772796602966e-05)
('Hong Zhang', 7.854095286461262e-05)
('Ajith Abraham', 7.845321957796962e-05)
('Mani B. Srivastava', 7.835090136234616e-05)
('Ying Liu', 7.826418075094073e-05)
('Luc J. Van Gool', 7.79756172258558e-05)
('Ying Chen', 7.78768512304891e-05)
('Tao Zhang', 7.721288842944203e-05)
('Donald F. Towsley', 7.712843687402171e-05)
('Jun Liu', 7.671917801147302e-05)
('C. C. Jay Kuo', 7.625128869758413e-05)

Figure 1

TOP 20: Edge Betweeness scores for dblp2005 graph WARNING: K WAS SET TO 5 DUE TO COMPUTATION TIME ISSUES.

(('Seiya Imoto', 'Tomoyuki Higuchi'), 9.351562587245186e-06)

(('Jianer Chen', 'Fenghui Zhang'), 9.351125398131143e-06)

(('Maria Bennani', 'Ahmed K. Elhakeem'), 9.350761073869435e-06)

(('Ahmed K. Elhakeem', 'Xiaofeng Wang'), 8.941100322039155e-06)

(('Pankaj K. Agarwal', 'Robert-Paul Berretty'), 8.925075253231555e-06)

(('Seiya Imoto', 'Minoru Kanehisa'), 7.530155073144892e-06)

(('Eiji Kamioka', 'Shigeki Yamada'), 7.132266678753389e-06)

(('Binxing Fang', 'Xiaofeng Wang'), 5.715636678008359e-06)

(('Chao Zhang', 'Shigeki Yamada'), 5.587445522352392e-06)

(('Jianer Chen', 'Wei Zhao'), 5.1699715295760265e-06)

(('Minoru Kanehisa', 'Jean-Philippe Vert'), 4.57744563781226e-06)

(('Jean-Philippe Vert', 'William Stafford Noble'), 4.28699956733186e-06)

(('Qiang Yang 0001', 'William Stafford Noble'), 3.4288448032873484e-06)

(('Jianer Chen', 'Guojun Wang'), 2.3226811046562666e-06)

(('Minoru Kanehisa', 'Tatsuya Akutsu'), 2.297081730846349e-06)

(('Lei Li', 'Binxing Fang'), 1.999989893480643e-06)

(('Yang Cui', 'Eiji Kamioka'), 1.9165388012797735e-06)

(('Seiya Imoto', 'Hideo Bannai'), 1.7730475038706684e-06)

(('Hideki Imai', 'Yang Cui'), 1.7145762570212963e-06)

(('Pankaj K. Agarwal', 'Jun Yang 0001'), 1.6440532666716433e-06)

Figure 2

dblp2006 PageRank and Edge Betweeness scores

TOP 50: PageRank scores for dblp2006 graph.

('Wen Gao', 0.0001240131307163 9335)
('Hui Li', 0.0001233023078582666)
('Chin-Chen Chang', 0.00012155789113614978)
('Xin Li', 0.0001181604234 2745631)
('Yong Zhang', 0.0001143310055 1674944)
('Wei Wang', 0.0001141839306766579)
('Qing Li', 0.0001137078990 2479058)
('Yan Zhang', 0.0001071760748 6426268)
('Wei Liu', 0.0001071526598924 3889)
('Wei Zhang', 0.0001068405 1247426157)
('Li Zhang', 9.897328971913328e-05)
('Yu Zhang', 9.855469285395953e-05)
('Ying Li', 9.544758462291256e-05)
('Jing Li', 9.488459658482037e-05)
('Xiaodong Wang', 9.468680055242117e-05)
('Yan Li', 9.368813789090804e-05)
('Wei Li', 9.254352922287345e-05)
('Ying Liu', 8.948725663577392e-05)
('Yong Yu', 8.929031168044768e-05)
('Francky Catthoor', 8.892675086632877e-05)
('Jian Li', 8.725189260558146e-05)
('Wei-Ying Ma', 8.717464224186699e-05)
('Hai Jin', 8.666108402959188e-05)
('Jian Wang', 8.523852018029665e-05)
('Laurence Tianruo Yang', 8.413681197740771e-05)
('Samuel Pierre', 8.38788358404156e-05)
('Lei Li', 8.327246374174167e-05)
('Hao Wang', 8.259999955237505e-05)
('Yang Yang', 8.248247554430839e-05)
('Xin Chen', 8.14946498437183e-05)
('Fang Liu', 8.141584215923699e-05)
('Lei Zhang', 8.126625653572787e-05)
('Masayuki Murata', 8.110350909961693e-05)
('Tao Zhang', 8.043902864329784e-05)
('Gang Wang', 8.018410512901535e-05)
('Hong Zhang', 7.973926909893576e-05)
('Yong Wang', 7.967609034166209e-05)
('Bin Li', 7.808916142631805e-05)
('Wei Zhao', 7.69881259922865e-05)
('Jing Wang', 7.671448876053606e-05)
('Gang Li', 7.656909818744382e-05)
('Tao Li', 7.635290669780309e-05)
('Mario Gerla', 7.552434294885482e-05)
('Liang Zhang', 7.516285920433627e-05)
('Jun Zhang', 7.46937833275883e-05)
('Wei Xu', 7.454573262707774e-05)
('C. C. Jay Kuo', 7.357036275218519e-05)
('Luc J. Van Gool', 7.233009163938589e-05)
('Elisa Bertino', 7.17407661522738 5e-05)
('Yan Chen', 7.146691397294093e-05)

Figure 3

TOP 20: Edge Betweeness scores for dblp2006 graph WARNING: K WAS SET TO 5 DUE TO COMPUTATION TIME ISSUES.

(('Andrew Hanushevsky', 'David Leith'), 8.076892008792238e-06)

(('Zhong Li', 'Zhihong Mao'), 5.024262118723 67e-06)

(('Paul Ammann', 'Jeff Offutt'), 4.649352377345648e-06)

(('Andrew Hanushevsky', 'Alexander S. Szalay'), 4.342361589210048e-06)

(('Geoffrey C. Bowker', 'William Turner'), 4.150853278734255e-06)

(('Les Gasser', 'William Turner'), 3.888210096544088e-06)

(('Jun Wang', 'Les Gasser'), 3.6635903689037534e-06)

(('Sushil Jajodia', 'Paul Ammann'), 2.94599791630748e-06)

(('Jeff Offutt', 'Robert M. Hierons'), 2.7258347913280782e-06)

(('Fouad A. Tobagi', 'David Famolari'), 2.3553354984242504e-06)

(('Lizhuang Ma', 'Zhihong Mao'), 2.308778514408332e-06)

(('Ian T. Foster', 'Alexander S. Szalay'), 2.0995572743980034e-06)

(('Huan Liu', 'Fouad A. Tobagi'), 1.8458861284342353e-06)

(('Michael J. Franklin', 'Ghaleb Abdulla'), 1.6827792399982846e-06)

(('Andrew Hanushevsky', 'Ghaleb Abdulla'), 1.6755487501843653e-06)

(('Eduard H. Hovy', 'Geoffrey C. Bowker'), 1.6655713566290503e-06)

(('Zhong Li', 'Jing Liu'), 1.6143831470726982e-06)

(('Gang Wang', 'Zhong Li'), 1.6059978390889248e-06)

(('Yan Zhang', 'Yuji Oie'), 1.569306934758252e-06)

(('Andrew Hanushevsky', 'Ani Thakar'), 1.294929723157863e-06)

Figure 4

dblp2005w PageRank and Edge Betweeness scores

TOP50: PageRank scores for dblp2005w graph.

('Wen Gao', 0.00017542500962232276)
('Chin-Chen Chang', 0.0001516923685712731)
('Wei-Ying Ma', 0.00013496270946328825)
('Xin Li', 0.0001276741456190978)
('Licheng Jiao', 0.00012409820637811528)
('Francky Catthoor', 0.00012203863886629683)
('H. Vincent Poor', 0.00011864045037993307)
('Zhaohui Wu', 0.00011773376451276849)
('Hans-Peter Seidel', 0.00011673838658232994)
('Xianlong Hong', 0.00011372875348091101)
('Mario Piattini', 0.00010882833957300277)
('Wei Zhang', 0.00010786463995350347)
('Tharam S. Dillon', 0.00010781923811267757)
('Minglu Li', 0.00010768188589323187)
('Samuel Pierre', 0.0001039655314442308)
('Witold Pedrycz', 0.00010159445135266541)
('Guanrong Chen', 0.00010107457723169963)
('Wei Liu', 0.00010051830339757158)
('Hai Jin', 9.961925295023577e-05)
('David J. Evans', 9.822363211006028e-05)
('Mario Gerla', 9.726039688501 41e-05)
('Bin Wang', 9.711535643782988e-05)
('Wei Wang', 9.709528216965254e-05)
('Alberto L. Sangiovanni-Vincentelli', 9.578750689175139e-05)
('Jinxiang Dong', 9.511909754286359e-05)
('Kaushik Roy', 9.345487238418834e-05)
('Heung-Yeung Shum', 9.270821429321069e-05)
('Hong Shen', 9.231102805061214e-05)
('Donald F. Towsley', 9.20039351354226e-05)
('Yong Zhang', 9.053833770711201e-05)
('Norman C. Beaulieu', 9.002159341561066e-05)
('Mahmut T. Kandemir', 8.996536817072494e-05)
('Philip S. Yu', 8.992507300252977e-05)
('Azzedine Boukerche', 8.933477999468094e-05)
('Fang Liu', 8.925815517962777e-05)
('Mani B. Srivastava', 8.910878779941838e-05)
('Ajith Abraham', 8.86335422475728e-05)
('Yvon Savaria', 8.840407055017919e-05)
('Tak-Wai Chan', 8.83862811438535e-05)
('Elisa Bertino', 8.728486393662993e-05)
('Yan Li', 8.727019103745124e-05)
('Jing Li', 8.700450401104181e-05)
('Yan Zhang', 8.691057882308394e-05)
('Xiang Li', 8.645051719824402e-05)
('Jiannong Cao', 8.598806490920331e-05)
('Lei Li', 8.555454806100373e-05)
('Tao Zhang', 8.554798500111323e-05)
('Chao-Tung Yang', 8.520155293730393e-05)
('Hsinchun Chen', 8.479429946140938e-05)
('Katsumi Tanaka', 8.444588662514062e-05)

Figure 5

TOP20: Edge Betweeness scores for dblp2005w graph WARNING: K WAS SET TO 5 DUE TO COMPUTATION TIME ISSUES.

(('Abdellah Salhi', 'Qingfu Zhang'), 9.354972662334738e-06)

(('José Rodríguez', 'Abdellah Salhi'), 9.354535473220693e-06)

(('Andrés Iglesias', 'José Rodríguez'), 9.351562587245186e-06)

(('Andrés Iglesias', 'R. Ipanaqué'), 9.351125398131141e-06)

(('Chai Quek', 'Ghee Ming Goh'), 9.351125398131141e-06)

(('John Guckenheimer', 'Warren Weckesser'), 9.35103796030833e-06)

(('Stefan Götz', 'Manuel Talón'), 9.155651606660876e-06)

(('Arvind Krishnamurthy', 'Robert Preis'), 8.959759451583032e-06)

(('Jean-Pierre Briot', 'Jacques Malenfant'), 8.784170614984413e-06)

(('Klaus Wehrle', 'Stefan Götz'), 7.954581662714507e-06)

(('Daming Shi', 'Chai Quek'), 7.553589056006919e-06)

(('Christian Queinnec', 'Luc Moreau'), 7.057190353150671e-06)

(('Jean-Pierre Briot', 'Christian Queinnec'), 7.056578288391008e-06)

(('Yaochu Jin', 'Qingfu Zhang'), 5.154314557737189e-06)

(('Oliver Junge', 'John Guckenheimer'), 4.797851926362987e-06)

(('Michael Dellnitz', 'Robert Preis'), 4.478100608980195e-06)

(('Robert Preis', 'Oliver Junge'), 4.477549264930815e-06)

(('Michael Dellnitz', 'John Guckenheimer'), 4.476750342305371e-06)

(('Daniel S. Yeung', 'Daming Shi'), 4.075222150206851e-06)

(('Jianyong Sun', 'Qingfu Zhang'), 3.92288749229387e-06)

Figure 6

The PageRank scores seem to change as the graph changes but the edge etweeness seems to have th same hierarchy. An example of this would be 'Wen Gao' being the the top edge betweeness for all three graphs and also 'Xin Li' is also in the top 5 for all three edge betweeness calculations.

PART C. LINK PREDICTION IN GRAPHS

i. based on dblp2005 create a graph dblp2005-core that includes nodes with degree d>=3.

I removed all of the nodes that had a degree less than 3 from the Giant Connected Component of the graph that had all of the co-authorship's from the year 2005 in it.
Before
Number of nodes: 106943
Number of edges: 300043
After
Number of nodes: 77153
Number of edges: 255815

ii. based on dblp2006 create a graph dblp2006-core that includes nodes with degree d>=3.

I removed all of the nodes that had a degree less than 3 from the Giant Connected Component of the graph that had all of the co-authorship's from the year 2006 in it.
Before
Number of nodes: 123808
Number of edges: 356968
After
Number of nodes: 112664
Number of edges: 317447

iii. compute the list of friends-of-friends FoF in dblp2005-core; this is the list of pairs of nodes that are exactly two-hops away in the network. FoF is the list of candidate edges to consider for the prediction problem

I selected all the node pairs that had a shortest path of at least 2 then I removed node pairs that had a distance of less than 2 between them. All of the left over pairs where stored in a variable called fofs

iv. compute the set of edges T that do not exist in dblp2005-core but exist in dblp2006-core. This is the set of target edges that we would ideally be able to predict.

I placed all of the edges that weren't in 2005 but were in 2006 in a variable called t1.

v and vi: compute the set of predicted edges P according to the following link prediction methods and compute the precision at k evaluation metric for values of k=10, 20, 50, 100, |T|, denoted as P@10, P@20, P@50, P@100, P@T for each method:

a. RD: random predictor

. P@10 is 0.0
P@20 is 0.0
P@50 is 0.1
P@100 is 0.05
P@T is 0.011169539164813442

b. CN: common neighbors

. P@10 is 0.0
P@20 is 0.0
P@50 is 0.0
P@100 is 0.02
P@T is 0.02558608228317272

c. JC: jaccard coefficient

. P@10 is 0.1
P@20 is 0.05
P@50 is 0.04
P@100 is 0.02
P@T is 0.017327650111193478

d. PA: preferential attachment

. P@10 is 0.0
P@20 is 0.0
P@50 is 0.08
P@100 is 0.04
P@T is 0.009181183592784779

e. AA: adamic/adar

. P@10 is 0.0
P@20 is 0.0
P@50 is 0.0
P@100 is 0.04
P@T is 0.026647825549789967

In this Section the highest score for P@T was the adamic/adar prediction. Most of the Prediction techniques did not manage to find anything in the top 10's or 20's but the Jaccard coefficient was able to find some values in the top 10's or 20's. The Jaccard Coefficient was strangely one of the lowest scores when it came to the measurement of P@T even lower than random predictor.

PART D. COMMUNITY DETECTION IN GRAPHS

WARNING: d WAS SET TO BE GREATER THAN 30 DUE TO COMPUTATION TIME ISSUES.

Community Sizes:
276
151
106
75
60
54
41
25
19
9

As can be seen the largest community is 276 in size and the smallest community is only 9 in size. This was obtained with k=10. this was also obtained where the set of nodes all had degree >30. This was due to computation errors.