Question 1

1. Degree centrality measures the number of outgoing edges from the node while eigenvector centrality measures the number of incoming edges to the node. Both of them are ways to measure the power of a node's influence inside the network, and PageRank and HITs use both of them for calculating prestige scores.

   PageRank assigns a small probability alpha by the programmer as the weight of incoming links and a large probability (1-alpha) as the weight of outcome links deciding the final prestige score. The incoming number links is decided by random jumping and is calculated only once. Similar to PageRank, HITs calculated both the score for outgoing links, authority score, and the score for incoming links, hub score. However, HITs use iterations to calculate the number of incoming edges, by starting from a small set, and updates the number of incoming edges while expanding the set. Thus, in HITs, the authority score and hub score depend on and are calculated from each other.

2. Edos-Renyi random graph model, Watts-Strogatz small world model, and Barabasi-Albert scale-free network model are all trying to build a network distribution model that are close to networks in the real word for better study of the network. Edos-Renyi random graph model is built by randomly assigning edges to a set of N vertices where N is large and edges assignment are independent form each other with probability p. This model is good for large networks with bounded degree; it is sharply concentrated with small diameter; it shows the Possion distribution. As an improvement of Edos-Renyi random graph model, Watts-Strogatz small world model not only has small diameter, but also high clustering coefficient. Barabasi-Albert Scale-Free Model is a better modeling than the previous two in terms of modeling Internet network and many other social phenomenon. Besides small diameter high clustering coefficient, scale-free model is able to show a long-tail property. Thus, its distribution follows a power law which states that for a node has degree k, the probability is k^(-gama), where gama is different for different types of network. Such model has other benefits besides the ones mentioned earlier: the log-log plot of it shows a straight line, and it is robust for normal nodes failure with an exception that it is vulnerable for the hub nodes failure.

   Barabasi-Albert scale-free network model uses the preferential attachment, saying that edges from the new vertex are more likely to link to nodes with higher degrees. That means, if a node already has high degree, this node is more likely to receive more incoming edges when new nodes are adding to the model and vice versa. When the model becomes static after adding infinite number of nodes into the model, the max number of incoming edges that a node may have is close to infinite, and the smallest number of incoming edges that a node may have is close to zero. As most incoming edges are gathered to the most highest degree nodes, there will be small amount of possible incoming edges distribute to those with low degree, and thus, the number of low degree nodes should be extremely large while the number of high degree nodes should be extremely small. The number of nodes with a certain degree should decrease exponentially as the degree increases. Thus, this model follows the power-law distribution.

Question 2

1. Ranking across different clusters may not be as meaningful as ranking inside clusters. Clustering helps in obtaining better ranking. Better ranking gives better clustering results as highly ranked nodes have larger impact. Thus, using ranking and clustering to enhance each other iteratively will improve clustering and ranking in each iteration. Such iterative enhancement in gives better clustering than simply clustering after ranking without iterative calculations as in SimRank-based clustering.

   RankClus's clustering and ranking are mutually enhanced. The ranking score are propagated in the network over different types of edges. The clusters are randomly partitioned in each iteration. The ranking propagation and clustering in each iteration enhance each other. In contrast, SimRank-based clustering simply propagate ranking score and then conduct clustering based on this one ranking score. The ranking score will not be enhanced by the previous clustering result. As discussed above, a iterative ranking and clustering gives better ranking and clustering result. Thus, RankClus is more efficient than SimRank.
2. APA: author to paper to author meta-path. An author is similar to another author if they are co-authors in many papers.

   APVPA: author to paper to venue to paper to author. An author is similar to another author if they have published many papers in the same venue.

   APTPA: author to term to author meta-path. An author is similar to another author if they have published many papers that share same terms. In other words, the authors may work in the same topics.

   Authors clustered using APA are real world friends or share many common friends as they have worked together or have worked with same people.

   Authors clustered using APVPA may have similar influence in their working area. As some venues are highly ranked and some are not, authors who have published in the highly ranked venues are clustered with authors who have published in the venues that have similar influence.

   Authors clustered using APTPA may work in similar fields and directions. They use similar terms in their paper topics.

   To allow users to give guidance in meta-path choices, PathSelClus can be used. The user can give seeds to infer that some authors should be in one cluster or some authors should not be in one cluster. After using the each meta-path to generate a clustering result as usual, we give each clustering result a score. The more consistent the clustering is with the user guidance, the higher the clustering score is. Then, choose the meta-path that generates the highest clustering score.

Question 3

1. PubMed search engine classifies the querying key words to be the following queried types: Subheading, All Fields, MeSH Terms. Thus, we need to find meta-paths that links queried words to these three types.

   When a journal article is indexed, the following parameters are extracted: article Type, Secondary identifiers, Language, Country of the Journal or publication history.

   The network may have queried words, MeSH terms, Paper, Article Type, Secondary identifiers, Language and Country these types of nodes. Papers have links to article type, secondary identifiers, and language and country. Article type and secondary identifiers are linked to MeSHterms and queried words are classifies to some MeSH terms.

   Meta-paths example:

   (1) Queries --- MeSH terms --- article type --- paper. This meta-path says that if some word in a query is linked to a MeSH terms that is in a paper's article type, then this query is likely trying to ask for this paper.

   (2) Queries --- MeSH terms --- secondary identifiers --- paper --- secondary identifiers --- paper. This meta-path says that if some word in a query is linked to a MeSH terms that is in a paper's secondary identifiers, then this query is likely trying to ask for papers that share common secondary identifiers as that paper.

2. As PubMed has huge amount of data and only a small amount can be human-labeled, knowledge propagation is important for sparse network to collect enough information for classification. It is also important to select and only select the unambiguous ones as the initial training data. Moreover, different knowledge would not have equal weight in deciding the classification result. For papers that are popular used, widely cited, and are more related in research topics, the papers should be more likely to be found. Thus, ranking should be used when doing classification. Also, as ranking across different classes may not be as meaningful as ranking within class. Thus, ranking and classification should mutually enhance each other in an iterative way, and RankClass should be used.

   To improve RankClass, we can use user-guided seeds to give guides on which meta-path should be more likely to be considered to propagate knowledge for the classification. Also, as the data is sparse, we can also combine clustering within classification to obtain better ranking distribution for clusters within classification. Also, as the label purity is also a problem, we can first find some unambiguous words that can be substitute for ambiguous words, and put more weights on those unambiguous words for better clarify queried problems.

Question 4

PathSim

Top 10 using APVPA for Christos Faloutsos
Linli Xu
Bin Cao
Tao Wang
Srujana Merugu
Richard Nock
Xuanhui Wang
Martin Grohe
Oded Maron
Corin R. Anderson
Jeffrey C. Schlimmer

Top 10 using APTPA for Xifeng Yan
Alina Beygelzimer
Dushan Z. Badal
Devika Subramanian
David K. Jefferson
Jayant Kalagnanam
Stefano Lonardi
Simeon J. Simoff
Lee-Feng Chien
Gio Wiederhold
A. J. Feelders

Top 10 using APVPA for AnHai Doan
Teddy Candale
Alina Beygelzimer
Xiaolei Li
Frederick Hayes-Roth
Jayant Kalagnanam
Gui-Rong Xue
Paolo Ferragina
Alberto Maria Segre
Simeon J. Simoff
Haiquan Li

Top 10 using APTPA for Jamie Callan
Lizhu Zhou
Wook-Shin Han
Srikanth Bellamkonda
Derek Long
Val Tannen
Stefano Spaccapietra
Shai Shalev-Shwartz
Devika Subramanian
Jayant Kalagnanam
Paolo Ferragina

P-PageRank

Top 10 using APVPA for Christos Faloutsos

Christos Faloutsos
Lizhu Zhou
Teddy Candale
Ian H. Witten
Takashi Okada
Linli Xu
Claudia Bauzer Medeiros
Srujana Merugu
Sachindra Joshi
Wook-Shin Han

Top 10 using APVPA for AnHai Doan

AnHai Doan
Lizhu Zhou
Teddy Candale
Ian H. Witten
Takashi Okada
Linli Xu
Claudia Bauzer Medeiros
Srujana Merugu
Sachindra Joshi
Wook-Shin Han

Top 10 using APTPA for Xifeng Yan

Clement T. Yu
David J. DeWitt
Elke A. Rundensteiner
Zheng Chen
Georges Gardarin
Rajeev Rastogi
Charu C. Aggarwal
Dimitris Papadias
Gio Wiederhold
Soumen Chakrabarti_

Top 10 using APTPA for Jamie Callan

Clement T. Yu
David J. DeWitt
Elke A. Rundensteiner
Zheng Chen
Georges Gardarin
Rajeev Rastogi
Charu C. Aggarwal
Dimitris Papadias
Soumen Chakrabarti
Gio Wiederhold