An Empirical Study on IMDb and its Communities Based on the Network of Co-Reviewers

Maryam Fatemi
Queen Mary University of London
Maryam.Fatemi@eecs.qmul.ac.uk

Laurissa Tokarchuk
Queen Mary University of London
Laurissa.Tokarchuk@eecs.qmul.ac.uk

Abstract

The advent of business oriented and social networking sites on the Internet have seen a huge increase in number of people using them in recent years. With the expansion of Web 2.0, new types of websites have emerged such as online social networks, blogs and wikis. Their popularity has resulted in exponential growth of information on the web and interactions overload thus making it harder to access useful or relevant information. Recommender systems are one of the applications employed to address this problem by filtering relevant information and enhancing user experience. They traditionally use either the content of items of the websites (content-filtering recommender systems) or the collaboration between the users and items such as rating (collaborative-filtering recommender systems) or a combination of them (hybrid recommender systems). However due to the nature of data they use, they all have one or more weaknesses such as cold start, sparsity of data, scalability problems and overspecialised recommendation. Social networks and other similar websites have new types of data which can be used in recommender systems thus have the potential to overcome these shortcomings. However without a good understanding of the properties and structure of these online social websites, the applications can not be accurate. This paper presents an empirical measurement study of the properties and structure of one such social websites. It examines an online movie database, and the interactions between reviewers and attempts to construct a social network graph based on the network of reviewers. The resulting network is confirmed as the power-law, small-world and scale-free. It identifies the highly connected clusters and shows that the content of these subgroups are diversified and not limited to similar tags. Finally the implication of these finding is discussed in order to enhance current recommender systems enabling them to provide diverse results while overcome their shortcomings.

Categories and Subject Descriptors C.4 [Computer Systems Organization]: Performance of Systems|- Measurement techniques; H.3.5 [Information Storage and Retrieval]: Online Information Services| Web-based services; H.3.5 [Online Information Services]: Web-based services

General Terms Measurement, Experimentation

Keywords Web 2.0, Social Network, Community Detections

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MPM'12, April 10, 2012, Bern, Switzerland
Copyright 2012 ACM 978-1-4503-1163-2/12/04…$10.00.

1. Introduction

With the expansion of Web 2.0 new generation of services have rapidly emerged. Its characteristics of user-centred design, interoperability and information sharing has facilitated collaboration of users and provided a platform for rich user experience to publish dynamic content [3]. Consequently this advance in technology has led to creation of the new generation of services such as wikis, blogs, social networks and social media websites. These new types of services enable people to share data or interact socially with each others and form communities. They also allow them to create content in blogs, wikis and micro blogs and contribute to content generation by tagging, commenting or ratings. Websites such as Google+, Facebook, YouTube, Wikipedia, Twitter, Last.fm, Amazon, and Flickr are among some of the more successful websites exploiting these new concepts. The amount of information posted by end-users is growing exponentially and this has lead to information overload. Although the information available is invaluable, the sheer amount of data and its ambiguity due to various format of it makes the usage limited, inadequate, inefficient and confusing as specific data will be hard to retrieve. Recommender systems have been proposed to address this problem by filtering the relevant and useful data. They traditionally use either the properties of the content of items of the websites (content-filtering recommender systems) or the collaboration between the users and items such as rating (collaborative-filtering recommender systems) or a combination of them (hybrid recommender systems). However due to the nature of data they use, they all have one or more weaknesses such as cold start, sparsity of data, scalability problems and overspecialised recommendation. Social networking websites have new types of information available which could address the common shortcomings of the current recommender systems. Hence they could play an important role in the application development, information dissemination and knowledge discovery. Subsequently understanding the characteristics and structure of these online social websites is important to improve the current systems, design and deploy new applications to enhance user experience and increase revenues.

Social networks are founded on identification of relationship between members, groups and organisations. They are good at illustrating how members are linked together through various social connections, such as values, ideas, friendship, collaboration, interests and information sharing. Online social networks are now an accepted way for people to interact, express themselves and share information. Social Network Analysis (SNA) techniques study and analyse the network structure, its ties, actors and social relationships utilising mathematical tools and

graph concepts. This paper utilises SNA techniques to study and analyse the social network of IMDb[1] (Internet Movie Database). The results confirm the power-law, small-world and scale-free properties of its network. The social network analysed is constructed based on the network of reviewers for movies. The resulting network consists of densely connected nodes with high-degrees which indicate strongly connected sub-structures, called communities. These structures are important as nodes belonging to a community tend to have very similar properties. This is important for recommender systems as they could produce recommendation within certain categories of interest in each community. The results are more effective comparing to using tags or item similarities as they tend to be static information, not standardised and could be too specialised. This paper examines four well known methods to discover communities and evaluates in detail their usefulness. It also shows that the content of these communities are diversified and could be used by various applications to identify interests of users. Finally, the implications of these finding is discussed to improve the relevant services and applications on social networks domains, such as recommender systems.

The paper is organised as follows: Related works in social network analysis and community detection algorithms are provided in section 1.1 Building and analysing of the online social network of movies is discussed in section 2. It also discussed the structural properties and communities. Finally it concludes the findings and their implications in section 3.

1.1 Related Works

The shape of the network helps determine its usefulness to its members. Smaller, tighter networks can be less useful to their members but networks with lots of weak ties to outside, create opportunities for brokerage due to having access to a wider range of information. A deeper understanding of underlying topology of such networks is essential in order to have successful application of their data. This would not only ensure the robustness and security of online social networks, but also would help understanding their impact on the future Internet and applications. Several attempts have been made to visualise and analyse social networks structure. A recent study on friendship relationship between social network users was done by Mislove et al [4]. Several public profiles on social network websites of YouTube, Orkut, Flickr and LiveJournal were crawled and a large scale measurement on their structure was performed. The outcome of this research showed that online social networks should have the small- world (small-world networks have a small diameter and but high clustering), scale-free (indicating that high-degree nodes tend to connect to other high-degree nodes, and low degree nodes tend to connect to low-degree nodes) and power-law properties.

Communities or subgroups structures are widely defined as group of node which are more densely tied to one another than any other part of the network. Partitioning of nodes into subgroups is important in understanding the likely behaviour of the network as a whole and can provide insight into how network function and topology affect on each other [1]. There are mainly two approaches to identify subgroups; Bottom Up (agglomerative) or Top Down (division). The former focuses on identifying the larger structures out of smaller most similar components, whereas the latter looks at the whole network and recognises substructures as areas of the graph denser than other parts and separated from the rest. The main draw back of agglomerative methods is their tendency in finding the core of the communities and leaving the border nodes out. This is due to the stronger similarities between the core nodes than the peripheral nodes. Community detection methods are further classified into global or local community detection depending on whether they use the global information or local knowledge in calculating similarities.

A well known agglomerative algorithm is proposed by Palla et al [8], which finds overlapping communities and is based on clique-percolation. In the clique-percolation method a community is defined as a union of all *k*-cliques (complete sub-graphs of size *k*) that can be reached from each other through a series of adjacent *k*-cliques where two cliques are adjacent if they share *k-1* nodes. With the increase in *k*-cliques, communities become more connected but would shrink. One of the advantages of this method is that it allows researcher to analyse the network at a higher level of organisation and identify the communities that play important role within the web of communities.

Lancichinetti et al [7] agglomerative algorithm also finds overlapping and hierarchical communities. It is based on the local optimisation of a fitness function, which calculates the strength of community with a resolution parameter α. The fitness is based on the trade of between the internal and the total degree of the cluster. Nodes are randomly added/removed until a local maximum in fitness value is reached. The resolution parameter α creates the hierarchical level of network. For $\alpha<0.5$ only one community can be identified. Subsequently for $\alpha>2$ smallest community are identified. Due to the high computational complexity it is infeasible to use this method for larger networks, but for smaller networks the resulting clusters display a good natural hierarchical structure of the networks. The Palla et al method [8] identifies overlapping communities mostly corresponding to well-defined categories, where as Lancichinetti et al method [7] categories are more mixed.

Blondel et al [6] method called Louvain is a widely used method based on the modularity optimisation which can analyse significantly larger networks very accurately with considerable lesser computational time. First it finds small communities by optimising the modularity using local information and secondly it aggregates small communities. These steps are repeated until maximum modularity optimisation is achieved.

Modularity *Q* [10] is defined as the fraction of the edges in the network that are connected in the same communities minus the expected value of the same quantity of nodes in a network with the random connections between them. Values of *Q* approaching *1* indicate a network with strong community structure [9]. Equation (1) is the modularity for a network with n_c partitions, defined as n_c x n_c size matrix *e* where the elements e_{ij} represent the fraction of total edges starting at a node in partition *i* and ending at a node in partition *j*. Then, the sum of any row (or column) of *e*, a_i corresponds to the fraction of links connected to *i*.

$$Q = \sum_i \left(\left(e_{ii} \right) - a_i^2 \right) \quad (1)$$

Where $a_i = \sum_j e_{ij}$

De Meo et al [5] proposed CONCLUDE (COmplex Network CLUster DEtection) which is based on Louvain method. It is a relatively fast community detection algorithm which uses both global and local information. It first looks at the global information of the network to establish the edges that contribute to creation of the community structures by measuring the *k*-path edge centrality. *K-path* edge centrality is based on the propagation

[1] http://www.imdb.com/

	Movies	*Reviews*	*Edges*	*Reviewers*
All	*21,894*	*276,795*	*3,860,809*	*133,318*
NWO30	*576*	*22,709*	*9,049*	*14,527*

Table 1. Data Collected from IMDb

of messages by using simple random walks of length at most *k*. Then it ranks the edges using a random walker. In the next steps it would calculate the proximity (i.e., the inverse of the distance) between pairs of connected nodes. Finally it partitions the network based on Louvain method aiming to optimise the network modularity. Needless to say due to the ranking process, CONCLUDE is computationally efficient which makes it more worthy to study large social networks.

In this paper we will depict and analyse the network of movies on IMDb (Internet Movie Database) and its underlying community structure using the four discussed community detection algorithms. The study of IMDb is interesting because users review diverse topics that is interesting to them hence the communities of movies linked by their reviewers could capture diverse interests regardless of their genre tags.

2. Building and Analysing the Social Network of Movies

Information from IMDb website was collected using a crawler tool called Screen-Scraper[2] over a period of 3 weeks in November 2011. It collected all the data of the movies and their reviewers. In total *21,894* movies information and *276,795* reviews from *133,318* users were collected summarised in Table 1.

To create the social network of IMDb, movies are presented as nodes and are linked if at least one reviewer has commented on both of them. The edges are weighted by the number of common reviewers between a movie pair. For instance, if movie *A* and *B* have been commented by *6* unique reviewers, the weight of their edge is *6*. For a reviewer *R* who has reviewed a set of movies *M*, $R(M) = \{M_1, M_2, .., M_n\}$, there exists $n(n-1)/2$ possible links amongst the *n* movies. In total for the collected data of *21,894* movies *3,860,809* edges exist. The visualisation and some of the measurement was performed using Gephi[3], an interactive visualisation and exploration tool for networks and complex systems, dynamic and hierarchical graphs. Due to the large size of the network and computational limitations, the study presented in this paper is narrowed to a subset of the heavy reviewed movies; movies reviewed by more than 30 users (hereafter called NWO30). It captured movies with more interests from reviewers and with stronger ties between them, thus indicating more common properties and better structure to study. The eliminated nodes are outliers movies with less common reviewers and hence less popularity. However in future work, we aim to further study the whole network of movies without removing any outliers and examine their effect in the overall structure.

Graphs are usually laid out with force-based algorithms for better visualisation. Their theory is that linked nodes attract each other and non-linked nodes are pushed apart. The most efficient layouts algorithm is Force Atlas, which is used in this paper to show the network of movies. Nodes are also arranged according to their degree rank.

[2] http://www.screen-scraper.com/

[3] http://gephi.org/

2.1 Network Measurement

Many different measurements can be carried out on a network; but here we limit the measurement to a few relevant metrics to identify the influential characteristics of the network.

Social networks are usually described by power-law degree distributions [2, 4], $P(k) \sim k^{-\lambda}$ where *k* is the node degree and $\lambda<3$. Networks which follow the power law distribution share a common structure of a relatively small amount of nodes connected by large number of edges that are not random. Random networks generally follow the Poisson distribution [4]. Degree distribution for NWO30 was calculated and is shown in Figure 1. The graph follows approximately the power-law distribution with $\lambda=1.289$.

Other important measurements presented here are Average Path Length *l*, Clustering Coefficient *c* and the Network Diameter *d*. These three metrics are important in network analysis as they indicate whether a network has the small-world and scale-free phenomena. Small-world phenomenon is often associated with the famous six-degrees of separation. A graph is considered to be small-world if most of the nodes who are not neighbours can be reached from every other node by a small number of steps. Usually small-world networks have a small diameter and high clustering. For NWO30 the Average Path Length *l* is *2.379* and Network Diameter *d* is 6; considerably small for the size of the network with *576* nodes. Also the Clustering Coefficient *c* is *0.752* which is much higher than an equivalent random network confirming that NWO30 network is small-world. It also confirms that the network is scale-free. Scale-free networks are a group of power-law networks where the high-degree nodes tend to be connected to other high degree nodes. A high clustering coefficient also suggests that the network has tightly knit groups. This is expected, as users who reviewed and liked two movies have similar interest and thus the two movies have a great deal of similar features. In general the observed result suggests that the structure of the IMDb network is a social graph with number of small tightly clustered communities held together with nodes with high degree.

2.2 Community Structure

Many networks have tightly interconnected group of nodes called communities. Community identification is very important and useful because nodes belonging to the same community tend to have similar properties. This can help recommender system to produce diverse recommendation and restrict them within the categories in a community. In the next sections four well-known methods of community detection algorithms are applied on the NWO30 and their outcomes are compared and evaluated.

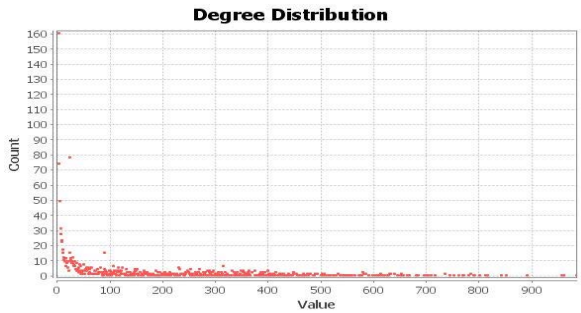

Figure 1. Degree Distribution for NWO30, following a power-law distribution with $\lambda= 1.289$.

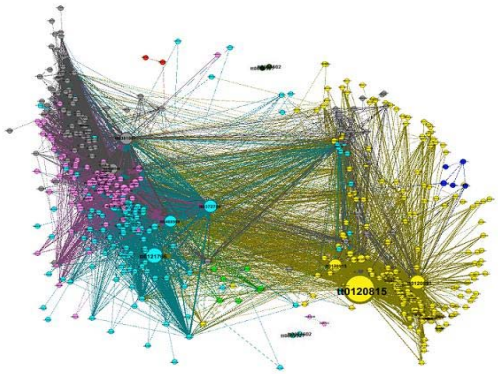

Figure 2. 12 Communities found by CONDLUDE.

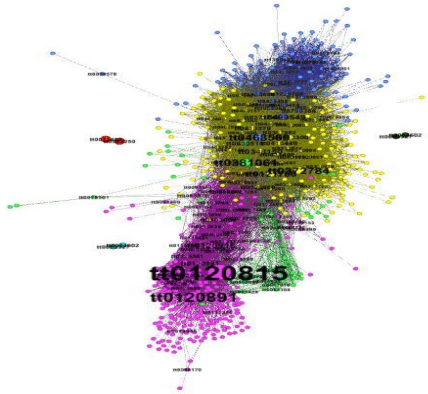

Figure 4. 7 communities detected by Louvain algorithm.

2.2.1 CONCLUDE Method

CONCLUDE [5] attempts to identify the communities based on modularity optimisation however uses *k*-paths edge centrality when ranking and partitioning the smaller communities. CONCLUDE was applied to NWO30 network and found *12* community structures with the maximum modularity of $Q = 0.489$. Figure 2. shows the communities discovered. The largest community had *231* movies and the smallest only *1*. The core movies in the largest community was "*Saving Private Ryan*"and "*Star Wars: Episode I - The Phantom Menace*" and the highest genres in this community were *Thriller*, *Adventure* and *Drama*.

2.2.2 Lancichinetti et al Method

Lancichinetti et al method [7] identifies overlapping and hierarchical communities by local optimisation of its fitness function. Communities are revealed by peaks in the fitness histogram. For NWO30 network this algorithm found *4* community structures with the highest value of fitness $f = 786$ and modularity of $Q = 0.226$, shown in Figure 3. The largest community had *363* movies and smallest had *5* members. The core two movies in the largest community were "*Casino Royale*" and "*Star Wars: Episode III - Revenge of the Sith*" with the highest genres of *Action*, *Thriller*, and *Adventure*.

2.2.3 Louvain Method

Louvain Method [6] finds small communities by optimising modularity in a local way. It then aggregates nodes of the same community and builds a new network out of communities. The steps are repeated iteratively until the maximum modularity is attained. The Louvain algorithm was applied to NWO30 and identified *7* communities with the maximum modularity of $Q = 0.324$, shown in Figure 4. The largest community had *247* movies and the smallest only *1* member. The top two core movies with the highest degree in the largest community were "*Saving Private Ryan*" and "*Wild Wild West*", respectively and the highest three genres in this community were *Thriller* followed by *Adventure* and *Action*.

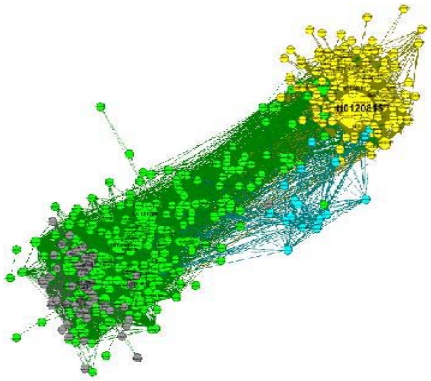

Figure 3. 4 communities identified by Lancichinetti.

2.2.4 Palla et al Method

Palla et al method [8] is based on clique percolation and *k*-cliques (clique is complete sub graph of *k* nodes) are rolled over the network through other cliques with $k - 1$ common nodes. The result is the huge number of various *k*-cliques communities. This method is implemented in CFinder application and is applied to NWO30. The maximum number of overlapping communities found were only *3* when $k=5$. Then the largest community had *173* members with *254* cliques of sizes *5* to *28* and the smallest community had *5* members with *1* clique of size 5. The maximum Modularity achieve when $k=5$ was $Q=0.298$. Figure 5. shows the *3* overlapping communities. The top two core movies with the highest degree in the largest community were "*The Dark Knight"* and "*Casino Royale*", respectively and the highest three genres in this community were *Thriller* followed by *Adventure* and *Sci-Fi*.

2.3 Evaluating Communities

Four well-known methods of community detection were applied on the network of the IMDb heavily co-reviewed movies by more than *30* people. The outcome of the community detection methods

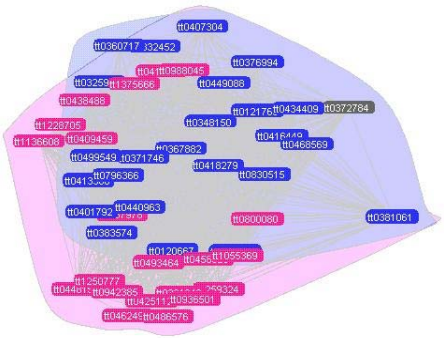

Figure 5. 3 overlapping communities identified by Palla when k=5, and the cliques inside them.

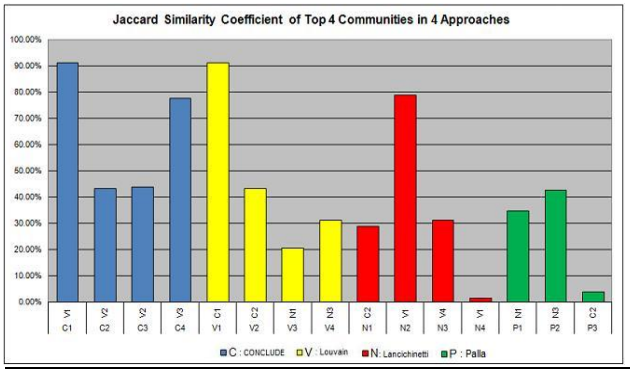

Figure 6. Jaccard Similarities Coefficient for the pairs of top 4 most similar communities identified by the four approaches.

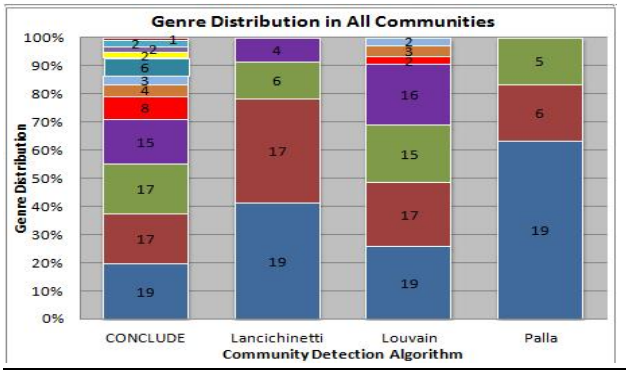

Figure 7. Genre Distribution for all communities identified using the four approaches.

were diverse and are evaluated here using various factors.

Modularity metric *Q* is widely used to show how well a network is partitioned. Subsequently the modularity is calculated for all four methods used here. The overall Modularity observed was not very high due to the randomness of the taste of the reviewers and the tight network, however CONCLUDE managed to attain the highest *Q* of *0.489*, followed by Louvain with *Q* of *0.324* indicating a more cohesive communities compared to Palla's and Lancichinetti of *0.298* and *0.226* respectively.

The similarities in the memberships of the four approaches are evaluated using Jaccard Similarity Coefficient [11]. Jaccard Similarity Coefficient is widely used to compare the similarity and diversity of the sample of data and is defined as the size of the intersection divided by the size of the union of the sample sets. Figure 6. shows the Jaccard Similarity Coefficient for the top four communities of the four the approaches of CONCLUDE, Lancichinetti, Louvain and Palla. Looking at the graph it is evident that CONCLUDE's communities have more similarities with the Louvain's communities while Lancichinetti's communities are more comparable with Louvain's. Although Louvain's communities have a relatively good similarities with Lancichinetti's communities, they still have more common members with CONCLUDE's communities.

Palla's underlying cliques in the communities however are difficult to compare with other three methods' due to the vast difference in number of cliques detected as Palla's method breaks the larger communities into smaller cliques of various seizes. It is however observed that Palla's *3* top communities with *k=5,* have more similarities with the Lancichinetti community *1* and *3*.

Next we examined the genre distribution of the communities in the four methods to determine how similar the groups of movies are and whether the movies are from diverse genres. Figure 7. shows the genre distribution in all the communities of CONCLUDE, Lancichinetti, Louvain and Palla. Cliques in Palla's largest community also had a good distribution of genres with average *8.708* genres per cliques (the smallest cliques of size 5 had movies with *4* different genres). It is observed that a good diversity of genres in the communities is of great use for recommender system to provide recommendation not limited to one category. Interestingly in all the four methods, their top two largest communities contained the same number of *19* and *17* different genres, except for Palla second largest community which had only *6* genres, even though their size differ. CONCLUDE, Lancichinetti, Louvain and Palla's two largest communities had *231*, *364*, *247,173* and *126, 204, 178, 12* movies, respectively. However CONCLUDE genres drops to *2*, *2* and *5* for the 3% of the movies in its smaller three communities. Similarly, the number of genres drops sharply in Lancichinetti communities to *6* and *4* for its remaining two communities. Louvain genre distribution however is consistent and only drops to *6* and *4* for the 1% remaining of movies in the two smallest communities.

Figure 8.a shows the genres distribution in the most similar communities and Figure 8.b. shows the distribution for the largest communities in all four approaches. The genre distribution is fairly similar for both graphs however the numbers of the movies in each community of the four methods are significantly different. As the distribution of the genre tags for the top communities were similar, two random communities with medium and small sets of movies where chosen for further analysis. Genres in Community *5* of CONCLUDE with *20* members were *Animation*, *Thriller*, *Mystery* and *Sci-Fi* in order of appearance. The smallest community in Lancichinetti method with 5 movies however had movies with genres of *Sci-Fi*, *Adventure*, *Thriller* and *Mystery*. Genres in community *4* of Louvain with *49* movies were *Adventure*, *Thriller*, *Crime* and *Sci-Fi* and in Palla's smallest community *3* with *5* movies were *Adventure, Sci-Fi, Fantasy* and *Drama.* Although the genres are different in each community, they still indicate a meaningful similarity between the movies. The smallest community in CONCLUDE had movies with one type of genre of *Action* where as Louvain's smallest has two genres of *Crime* and *Drama*.

Complexity and processing time of all the methods are also examined. As expected, CONCLUDE and Louvain method outperformed other methods by a huge difference. Their calculation time was a little less than two minutes where as Palla's was two hours on a normal desktop pc and Lancichinetti's was a little over two days on a Linux server. Computational complexity of Lacichinetti is $O(n^2)$ followed by Palla with $O(n^{ln(n)})$, CONCLUDE $O(k\ |E|)$ ($|E|$ number of edges) and Louvain with $O(n\ log\ n)$. Therefore when choosing a method an informed decision needs to be taken based on the trade-off between the size of the network and number of communities detected and whether a finer categories are desirable or broader communities.

NWO30	CONCLUDE	Lancichinetti	Louvain	Palla
Modularity	0.489	0.226	0.324	0.298
Complexity	O(k |E|)	$O(n^2)$	O(n log n).	$O(n^{ln(n)})$,
Time (hrs)	0.03	48	0.03	2
Communities	12	3	7	3
Ave Genres	8	11.5	10.5	10

Table 2. Summary of Community Detection Methods properties.

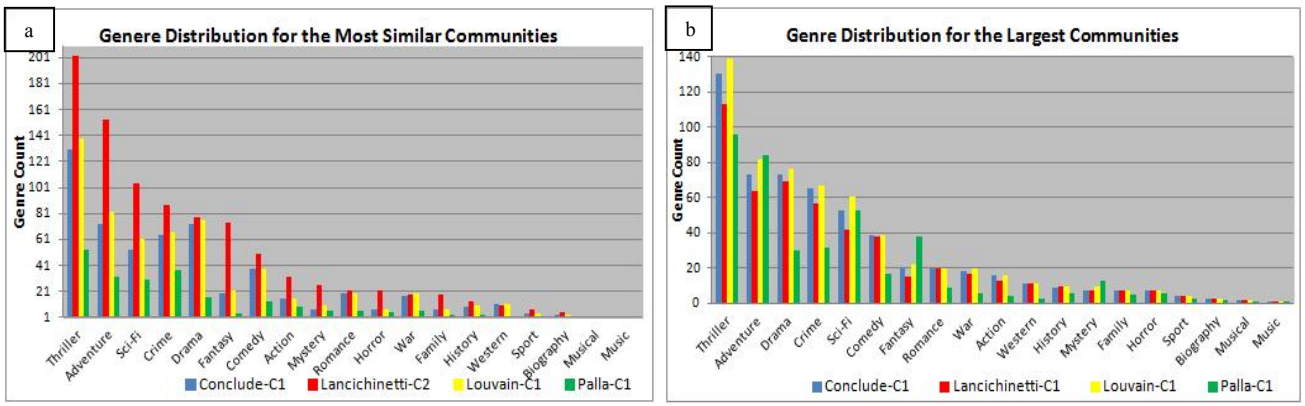

Figure 8.a, the genre distribution within the most similar communities in terms of their members. 8.b, the distribution of genre tags within the largest communities identified by four approaches.

In terms of efficiency, accuracy and meaningfulness CONCLUDE had the highest modularity, lower complexity and a better distribution of movies with different genres followed by Louvain. Palla et al communities however had finer cliques insides them which captured the sub genres for the movies better. Lancichinetti's communities were broader and more general. Table 2. Summarises the properties of the four community detection methods applied here and their results.

Further analysis of genres distribution of the movies is needed to better understand the correlation between the genre in community structure of IMDb with all movies and not the smaller subset.

3. Conclusion and Future works

In this paper the social network representation of the movies on IMDb is constructed where nodes represents movies and are only linked if a reviewer had commented on a pair of movies. The weight of each edge is calculated based on the number of common reviewers for the movies. The resulting network is an undirected, weighted graph. Further social network analysis shows that it follows the power-law distribution, has small-world properties, is scale-free and consists of tightly linked nodes. The network is then partitioned into smaller clusters of communities with shared or similar properties in movies. It also shows that the communities consist of diverse set of movies with different genres capturing diverse interest of the reviewers.

The outcome of this study is important for underlying applications such as recommender systems, which are trying to enhance user experiences by filtering data and personalising results. The network structure and community compositions have an impact on recommender system. It can help them to restrict the recommendations within certain categories given by the communities. This is more effective than trusting genres tag since they tend to have static information and generally not standardised. Using communities also enable recommender system to diversify their recommendation and not over specialised in one category. Also due to using new types of data (the interactions rather than item properties) theses systems can also overcome some of their weaknesses such as cold start and sparsity of data. They can also show the temporal changes in interests of the people due to the nature of social networks.

This is an ongoing research and the implication of these findings on underlying applications such as recommender system's design and performance will be investigated and the result will be published in future.

Acknowledgments

We thank EPSRC for its financial support and BT Research Centre for scientific collaboration.

References

[1] Pons, P. and Latapy, M. Computing Communities in Large Networks Using Random Walks. In ISCIS2005, 2005.

[2] Catanese, S., De Meo, P., Ferrara, E., Fiumara, G. and Provetti A. Crawling Facebook for Social Network Analysis Purposes. WIMS '11: Proceedings of the International Conference on Web Intelligence, Mining and Semantics, 2011.

[3] Anderson, P.: What is Web 2.0? Ideas, Technologies and Implications for Education. JISC Technology & Standards Watch, 2007.

[4] Mislove, A., Marcon, M., Gummadi, K., Druschel, P. and Bhattacharjee, B. Measurement and Analysis of Online Social Networks. Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, 2007.

[5] De Meo, P., Ferrara, E., Fiumara, G. and Provetti, A. Generalized Louvain Method for Community Detection in Large Networks- CONCLUDE. ISDA '11: Proceedings of the 11th International Conference on Intelligent Systems Design and Applications, 2011.

[6] Blondel, V. D., Guillaume, J., Lambiotte, R. and Lefebvre, E. Fast Unfolding of Communities in Large Networks- Louvain. In Journal of Statistical Mechanics: Theory and Experiment, P1000, 2008.

[7] Lancichinetti, A., Fortunato, S. and Kertész, J. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks. New Journal of Physics 11, 033015, 2009.

[8] Palla, G., Derenyi, I., Farkas, I. and Vicsek, T. Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature, 435:814, 2005.

[9] Bagrow, J. P. Evaluating Local Community Methods in Networks. Journal of Statistical Mechanics: Theory and Experiment, vol. 5, 2008.

[10] Girvan, M. and Newman, M. Finding and Evaluating Community Structure in Networks. Physical Review E, 69, 026113, 2004.

[11] Jaccard, P. Étude Comparative de la Distribution Florale Dans Une Portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles vol. 37, pp. 547–579, 1901.