Employee effort and productivity in organizations: A network data science approach

7/16/17

**Nan Wang; Evangelos Katsamakas**

# Abstract

Estimating the performance of employees is an important consideration in all organizations. This paper proposes a network data science approach to the estimation and visualization of employee effort, productivity and collaboration patterns. Using data from a software development organization, a bipartite weighted network is firstly constructed for developers’ commitment to project repositories. It is afterwards projected into two weighted one-mode networks: developer-to-developer and repository-to-repositories. For the former graph, two developers are defined to be connected if they once shared projects and for the latter, two repositories relate with each other if sharing developers. Techniques applied include graph theoretic metrics, power-law estimation, and community detection algorithms. We validate the existence of power-law relationships on project sizes (number of developers). We discuss implications for managers and future research directions. As a methodological contribution, the paper demonstrates how network data science can be used to derive a broad spectrum of insights about employee effort in organizations.

**Keywords:** networks, data science, network analysis, network science, ~~power laws,~~ employee effort, productivity, collaboration

# Introduction

Estimating the performance of employees is an important consideration in all organizations. This paper proposes a network data science approach to the estimation and visualization of employee effort, productivity and collaboration patterns.

While the methodology can be used in any type of organization, the data we use to illustrate an application of the methodology come from a software development organization. Software application development has turned into an enormously profitable business, with revenue from mobile application purely expected to exceed fifty billion USD by 2016 [1]. Accurate evaluation of developers’ achievement and contribution has been proved to be critical for the long-term development of companies as it is directly related to employee morale, overall productivity and creativity. Poor capability to discover and monitor coding performance prevents technical companies from transforming information on network activity and infrastructural capabilities into strategic knowledge [2, 3].

(REMOVE THIS PARAGRAPH FROM INTRO; ADD IT LATER IN PAPER IF NECESSARY) Although extensive historical study has applied social network analysis as a powerful solution to help understand organizational network performance and service interaction [4], most of them were designed in the context of binary networks. However, many real-life networks are widely recognized to be intrinsically weighted. Characterizing features and structures in weighted networks, therefore, has more important practical significance [5]. Bearing that in mind, this paper investigates a weighted bipartite network where developer is one mode and repository is another.

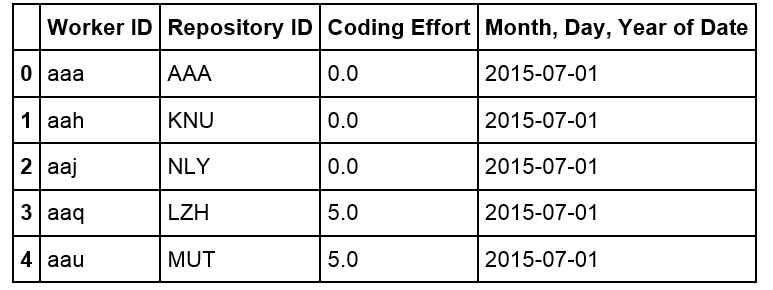
The questions we attempt to answer include identification of developers’ role, contribution, motivation, functionality and working preference, as well as in a broader sense, project membership, corporate cohesion, collaboration patterns and software development productivity.

We demonstrate how a network data science approach can help companies gain rich insights about these questions. We define network data science as the use of data science methods, tools and algorithms in the modeling and analysis of network (graph) data.

(WRITE A SUMMARY OF MAIN STEPS WE FOLLOW)

# Dataset and data preparation

The dataset used in this paper is the coding effort of 2621 developers to 1705 repositories in a real-life company in a period of 92 days, from July 1, 2015 to September 30, 2015. This panel data has four variables, namely date, Worker ID, coding effort, Repository ID. There are 172,354 records in total, where “Worker ID” and “Repository ID” identify each unique developer and project respectively. Data is gathered on past software development activity via source code repositories like Subversion and Git, and task tracking systems such as Jira. This dataset is kindly provided by the company BlueOptima. Table 1 shows a subset from the dataset.



**Table 1.** Sample of original dataset. Coding effort is a proprietary software measurement and is calculated through evaluating every change that software developers contribute to projects in terms of a series of metrics, such as volume, complexity and interrelatedness of codes (Newswire 2013)

There are many records with zero coding effort in the data. It is resulted from the fact that if a developer is involved in a project, coding effort will be recorded regardless of the absence of contribution. Therefore, to keep the information of developers’ involvement, we remain all data records.

# Modeling Effort and Contribution: Developers-Projects Graph

When relations are modeled between two different classes of objects, bipartite graphs arise naturally (Guillaume and Latapy 2004). A graph of customers and products, with an edge between a customer and a product, if the customer has bought that product, is a simple example of a bipartite graph.

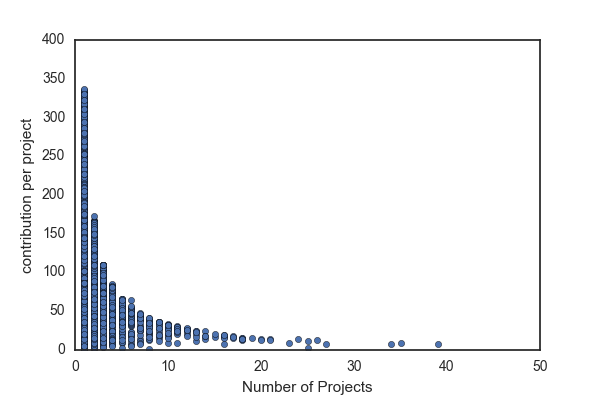
To construct the graph, we aggregate each developer’s contribution to every repository. The graph therefore stands for a static developer-to-repository network, where developers belong to one mode and repositories another. Edges represent developers’ contribution to projects and are weighted by coding effort.

* 1. **Graph Properties**

To understand the properties of this network, we apply a measurement called degree centrality to the graph. (Newman 2010)

Degree is the number of ties that a node has. In the developer-repository situation, nodes with degree of one can be interpreted as “focused developers” who work on only one project, or individual project which contain only one developer. It is found that during the entire period in concern, over 50% developers (1409 out of 2459) devote to only one project and above 40% projects (610 out of 1496) are individual projects. In addition, the maximum degree is 39 for all developer nodes and 96 for repository nodes, meaning that the most multi-tasking developer (Developer “ouj”) contributes to 39 projects and the largest project (Project “KGD”) is contributed by 96 developers. Also, the average degree is 2.3 for worker nodes and 3.8 for repository, meaning that, overall, each worker works on two projects and each repository contains four workers.

With most projects being small and a few large, it is reasonable to conclude that sizes of projects are polarized, indicating that the company’s software development business is a combination of a few principal projects and plenty of small projects. Moreover, if developers’ roles can be implied by the number of projects they work on and the contribution they make, for example, managers tend to distribute their contribution in a wider range of projects, we are probably able to see the workforce allocation. As it shown in Figure 1, developer nodes with large degree (many projects) but small-weighted edges (small contribution per project) can possibly be interpreted as project managers, while nodes with small degree (small project volume) and small-weighted edges (small contribution per project) can be considered as unproductive workers.



**Senior manager**

**Manager**

**Unproductive**

**Figure 1.** Implications of roles and productivity based on node degree and edge weight

* 1. **A Graph Recommender System**

Recommender systems have been created for various data science applications in a variety of areas (Ricci et al 2011). For example, Facebook utilizes recommender systems to suggest friends to users, iTunes and YouTube utilize similar machine learning and recommendation algorithms to suggest songs, videos and movies. Given this general theme, we attempted to create a recommender system to suggest the assignment of developers to projects.

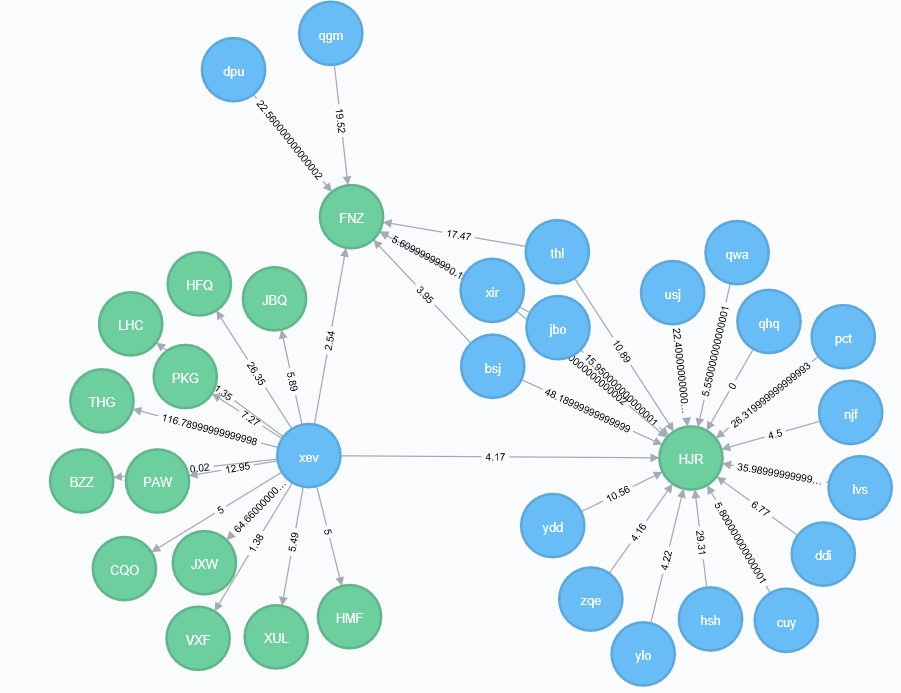
Collaborative filtering (CF), one of the most popular ways to implement recommender system, is a method of making [predictions](https://en.wikipedia.org/wiki/Prediction) or recommendations (filtering) about the interests of a user by collecting preferences or [taste](https://en.wikipedia.org/wiki/Taste_(sociology)) information from [many users](https://en.wikipedia.org/wiki/Crowdsourcing) (collaborating). There are three major types of collaborative filtering algorithms in the recommender system literature, namely memory based approaches (uses user rating data to compute the similarity between users or items), model based Approaches (uses [data mining](https://en.wikipedia.org/wiki/Data_mining), [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms to predict users' rating of unrated items), as well as hybrid approaches (combines the memory-based and the model-based algorithms). Each of those approaches has its own advantages and limitations (Ricci et al 2011; Breese et al 1998).

In the situation of developers’ contributing to projects, developers can be modeled as users in the recommender systems. Similarly, projects can be modeled as items and coding effort are ratings. The choice of modeling methods is usually affected by the data available and the purpose of recommender systems. Considering that our recommender system is to make recommendations rather than predictions, and content data (data describing users or items) is unavailable, we choose to apply memory based approaches.

Memory based approaches has advantages including content-independence, easy implementation and intuitive interpretation. (Breese et al 1998). However, it is sensitive and vulnerable to large sparsity data. Given the fact that in our case, there are 2621 users and 1705 items, and over half of users rated only one item, the data is expected to be sparse.

One of the common ways to address sparsity is to apply hybrid approaches. For example, use principle component analysis to compress a high dimensional user-item matrix containing abundant number of missing values into a much smaller matrix in lower-dimensional space. With the ability of effectively overcoming the problems of sparsity and loss of information, this method is widely applied in commercial recommender systems (Das et al 2007). However, it is complex and expensive to implement.

Therefore, instead of hybrid approaches, we turn to graph to address the problem of data sparsity. We firstly define that two users are neighbors if they share at least one item. In other words, two developers are neighbors if they work on at least one project. Instead of computing similarity of every two users, we only do calculations between neighbors.



**Figure 2.** ‘XEV’ is neighbored with both ‘dpu’ and ‘bsj’, by sharing the one project with the former and two with the latter.

To define the similarity of ui to u1, the following function is applied:

**Sim(u1, ui) =**

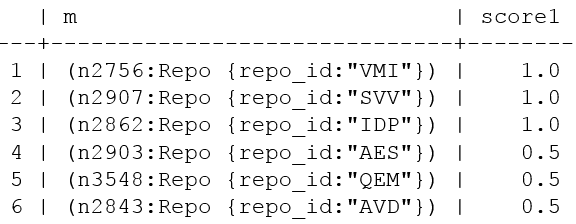
Number of Common Projects between u1, ui

Number of Projects of ui

The similarity of ui to u1 is the ratio of the number of projects they have in common over the number of projects ui works on. We further set a threshold of 0.5 for similarity, which means that developer ui is considered to be similar to developer u1 only if u1 work on at least half of ui’s projects.

We can afterwards make recommendations by answering either the question of what projects popularly shared by similar neighbors, or what projects actively contributed by similar neighbors.

Table 2 is a list of projects recommended by the system to the developer ‘xev’:



**Table 2.** Recommender system with similarity threshold being 0.5

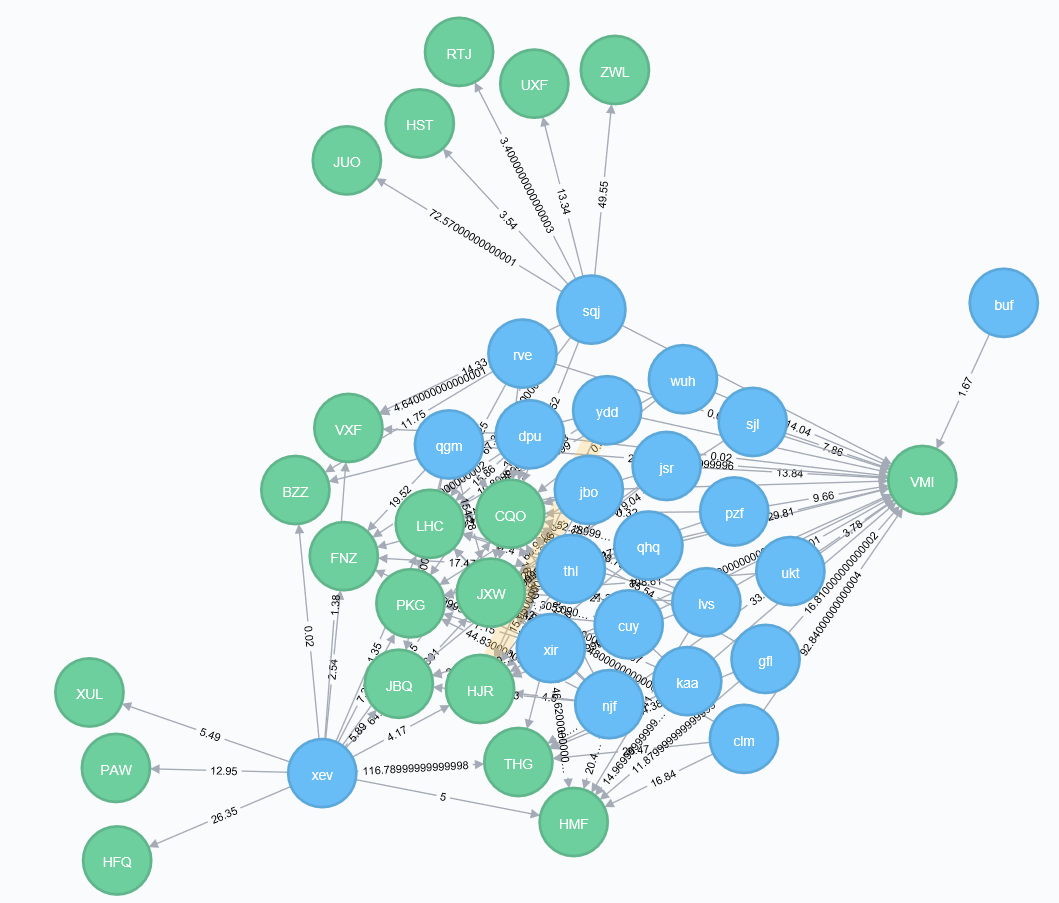
The scoring function calculates the ratio of similar neighbors who contribute to the recommended project.

**Score 1** **=**

Number of contributing similar neighbors

Number of similar neighbors

Assuming a manager is about to assign the developer ‘xev’ a familiar project, ‘VMI’, ‘SVV’ and ‘IDP’ would be worthy of consideration, given the fact that all xev’s nearest neighbors have worked on them, as it shown in figure 3.



‘xev’ individual projects

‘xev’ remote neighbors

‘xev’ similar neighbors

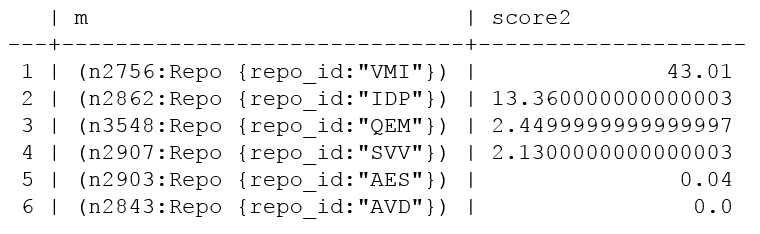
**Figure 3.** All xev’s nearest neighbors work on project ‘VMI’.

Table 3 is another list of projects recommended by the system to the developer ‘xev’. Different from the previous system, the scoring function has changed into the sum of similar neighbors’ contribution. Therefore, the system recommends new projects to the developer ‘xev’ in an order from ‘contributed the hardest by similar neighbors’ to ‘contributed the least by similar neighbors’.

**Score 2 =**

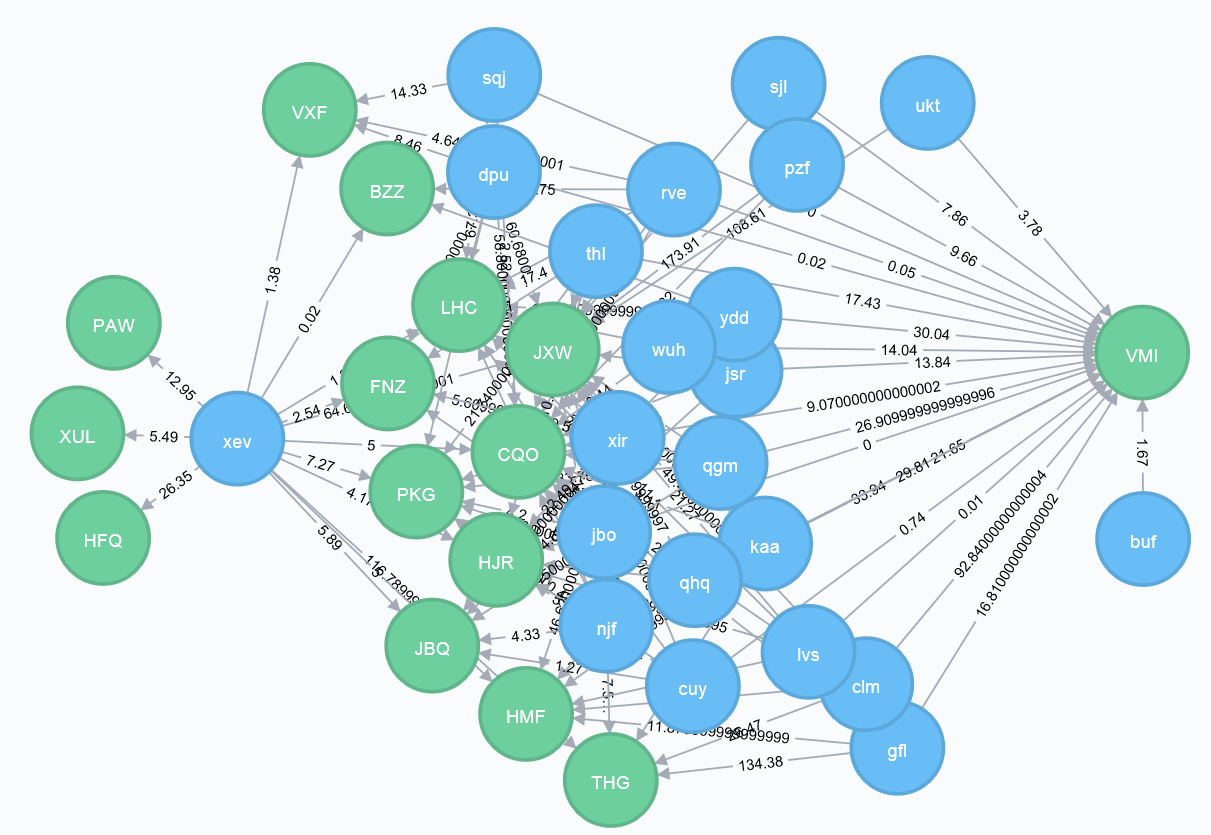
Sum (similar neighbors’ contribution)

If a manager wants to ensure developers’ contribution when assigning new projects, he probably prefers this recommendation scheme based on the assumption that a developer will contribute to a project in a comparable way as his similar developers. If nearest neighbors unanimously work hard on a project, the developer is likely to devote too.



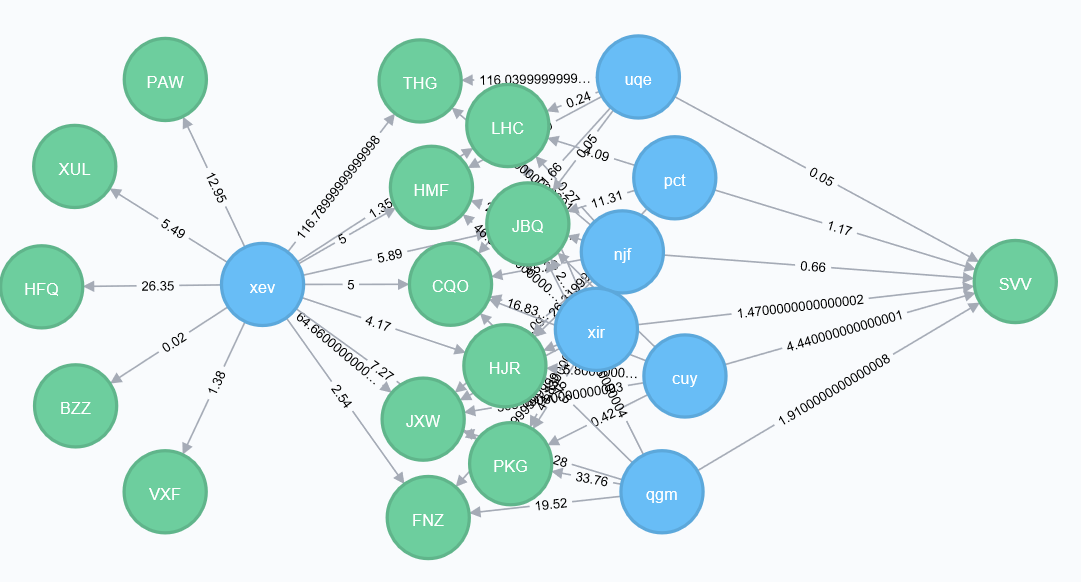
**Table 3.** AnotherRecommender system with similarity threshold being 0.5

More insights can be retrieved if the results of both recommender systems are combined. For example, although both project ‘VMI’ and ‘SVV’ are participated by all the nearest neighbors of developer ‘xev’, the former receives massive contribution (as it shown in figure 4) while the latter has little (as it shown in figure 5). Also, although half of nearest neighbors participate in the project ‘AVD’, none of them contribute anything in the concerning period.



Large contribution

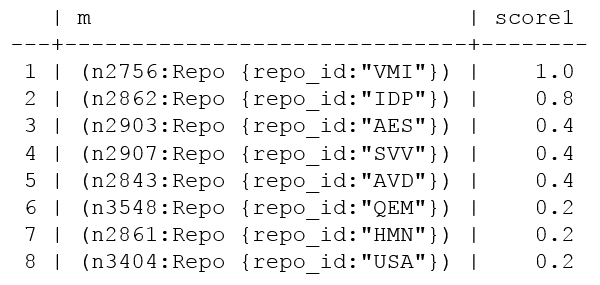
**Figure 4.** Most of xev’s nearest neighbors work hard on project ‘VMI’, especially developer ‘clm’ whose coding effort is about 93.



Small contribution

**Figure 5.** None of xev’s nearest neighbors work hard on project ‘VMI’.

Furthermore, we’ve found that different similarity threshold would affect recommendation results largely. For example, if we change the similarity threshold to 0.4, as shown in table 4, the results would be as follows and ‘VMI’ is the only project shared by all the developer’s nearest neighbors.



**Table 4.** AnotherRecommender system with similarity threshold being 0.4

An optimized graph-based recommendation system is beyond scope of this paper. Here we just implement a general-purpose way to demonstrate the simplicity and scalability of a graph-based approach. It doesn’t necessarily require user features or large volume of data for training. Moreover, it helps solve the data sparsity and computation problem of memory-based approaches.

* 1. **Conclusion**

With the weighted bipartite developer-project graph, we gain an overview of the structure and workforce allocation of the company’s software development business, as well as build an efficient graph-based recommender system which facilitates the project assignments of different purposes.

# Modeling Collaboration: Developers-Developers Graph

When relations among only one class of objects in a bipartite graph are needed, projection is usually applied to transform the graph from two-mode into one-mode (Zhou et 2007). The process works by selecting one set of nodes, and linking two nodes if they are connected to the same node of the other set. However, since one-mode projection is always less informative than the bipartite representation and the redistribution of weights would strongly affect the community structure, a proper weighting method is required to better retain the original information (Fan et al 2007).

In this paper, Newman’s weighted projection is applied, which adopts the weighting scheme presented by Newman. This method creates a one-mode network in which the out-strength of a node is equal to the sum of the weights attached to the ties in the two-mode network that originated from that node (Newman 2001).

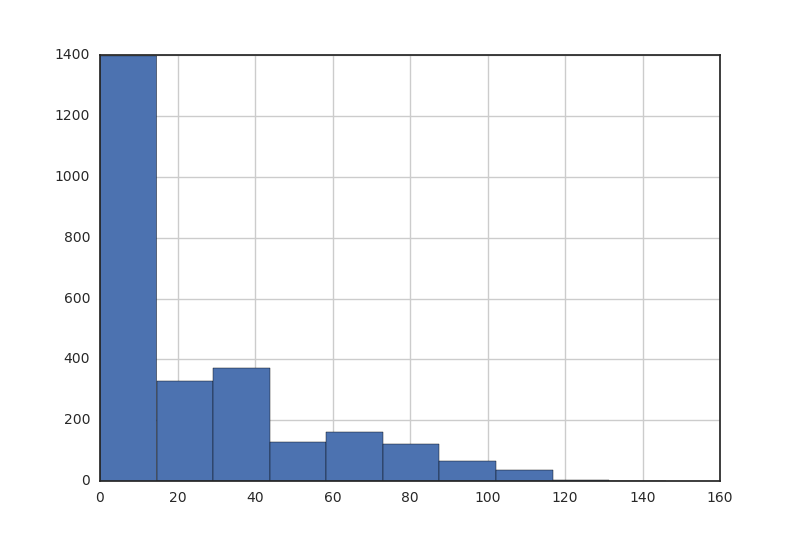
Accordingly, the developer-to-repository graph is projected into two separate one-mode graphs. One is a developer-to-developer graph with 2459 nodes and 61378 edges where each developer is a node and an edge exists between two nodes if both developers are on the same project, and another is a repository-to-repository graph with 1496 nodes and 16764 edges where repositories are nodes and edges represent shared developers.

* 1. **Graph Properties**

With the one-mode developer-developer graph, we can then elaborate on developers and their connections through studying properties of the developer-developer graph.

* + 1. **Degree Centrality**

We firstly apply degree centrality and plot out the distribution of degrees. Figure 6 shows that the degree distribution appears to follow a power-law distribution. (Clauset et al 2009). In the network theory, a network is named scale-free if its degree distribution follows a mathematical function called a power law. (Albert and Barabasi 2002). Unlike a random graph (i.e., new nodes attach to existing nodes with uniform probabilities), a scale-free graph displays preferential attachment of new nodes (i.e., some nodes have higher probability of attachment than others). In our developer-project case, it is intuitive that the network is not randomly formed. It is also in accordance with expectation that a developer, if having diversified skillsets or senior experience, is likely to collaborate with more other developers.



**Figure 6.** Node degree distribution for developer graph

In many real-world cases, the power-law behavior kicks in only above a threshold value (Xmin) for the input vector. To validate our assumption that nodes in the graph have preferential attachment, we apply a Statistical analysis called Kolmogorov-Smirnov test (Hazewinkel and Michiel 2001) and use the maximum likelihood principle. The algorithm attempts to find an optimized Xmin value for which the p-value of a Kolmogorov-Smirnov test is the largest (Clauset et al 2009).

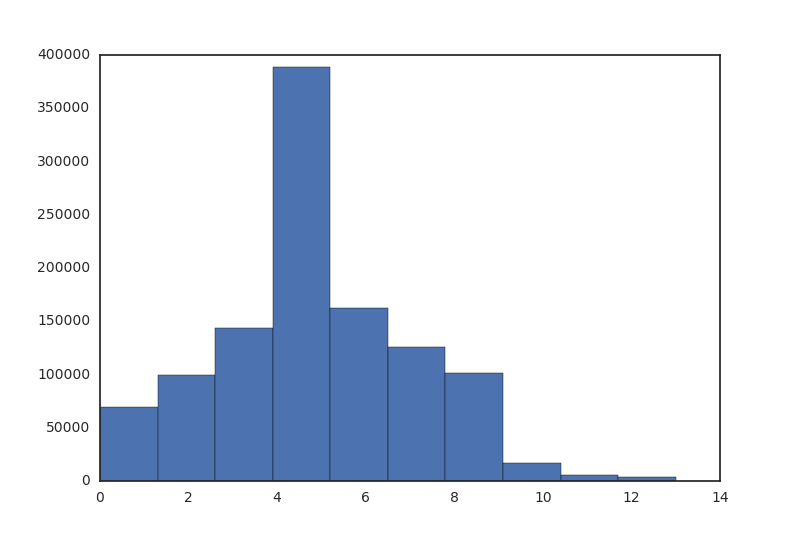
Test results are as follows. The combination of small test statistic (KS.stat) and bigger p-value (KS.p) together indicates a good fit of power-law distribution. As it shown in Table 5, KS.stat is 0.0346 and KS.p is 0.229. It therefore proves that the developer-to-developer network has the property of scale-free. (Marsaglia et al 2003)

|  |  |  |
| --- | --- | --- |
| **Measurement** | **Definition** | **Value** |
| Alpha | The exponent of the fitted power-law distribution. | 2.49 |
| xmin | The minimum value from which the power-law distribution was fitted. | 77 |
| logLik | The log-likelihood of the fitted parameters. | -5082.98 |
| KS.stat | The test statistic of a Kolmogorov-Smirnov test that compares the fitted distribution with the input vector. **Smaller scores denote better fit.** | 0.0346 |
| KS.p | The p-value of the Kolmogorov-Smirnov test. Small p-values (less than 0.05) indicate that the test rejected the hypothesis that the original data could have been drawn from the fitted power-law distribution. | 0.229 |

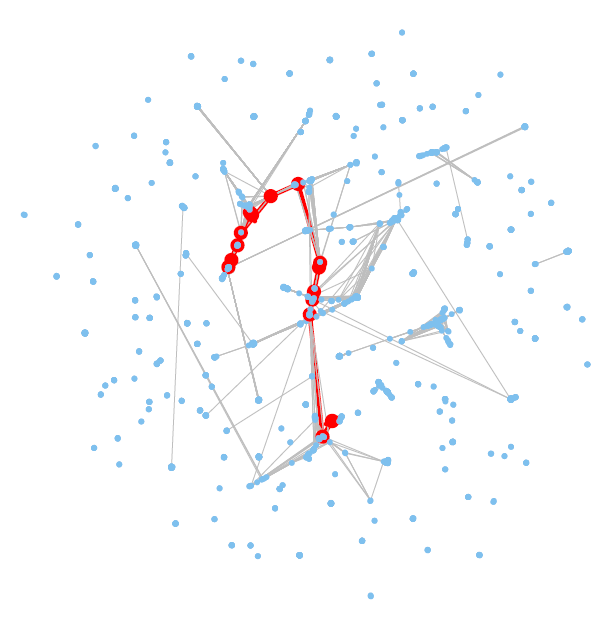
**Table 5.** Statistical tests for power-laws

* + 1. **Shortest Path Length and Closeness Centrality**

Shortest path length is the number of steps along the shortest paths for all possible pairs of network [nodes](https://en.wikipedia.org/wiki/Node_(networking)) (Zhan et al 1998). Figure 7 shows the distribution of shortest path length in the graph. Averaged at 4.86, length of the shortest path ranges from one to thirteen, meaning that a developer might expect to reach a randomly-selected developer in a typical distance of five, and the distance could be as small as one and as large as thirteen (as it shown in figure 8) in the company of 2,459 developers.



**Figure 7.** Distribution of Shortest Paths in the developers’ network

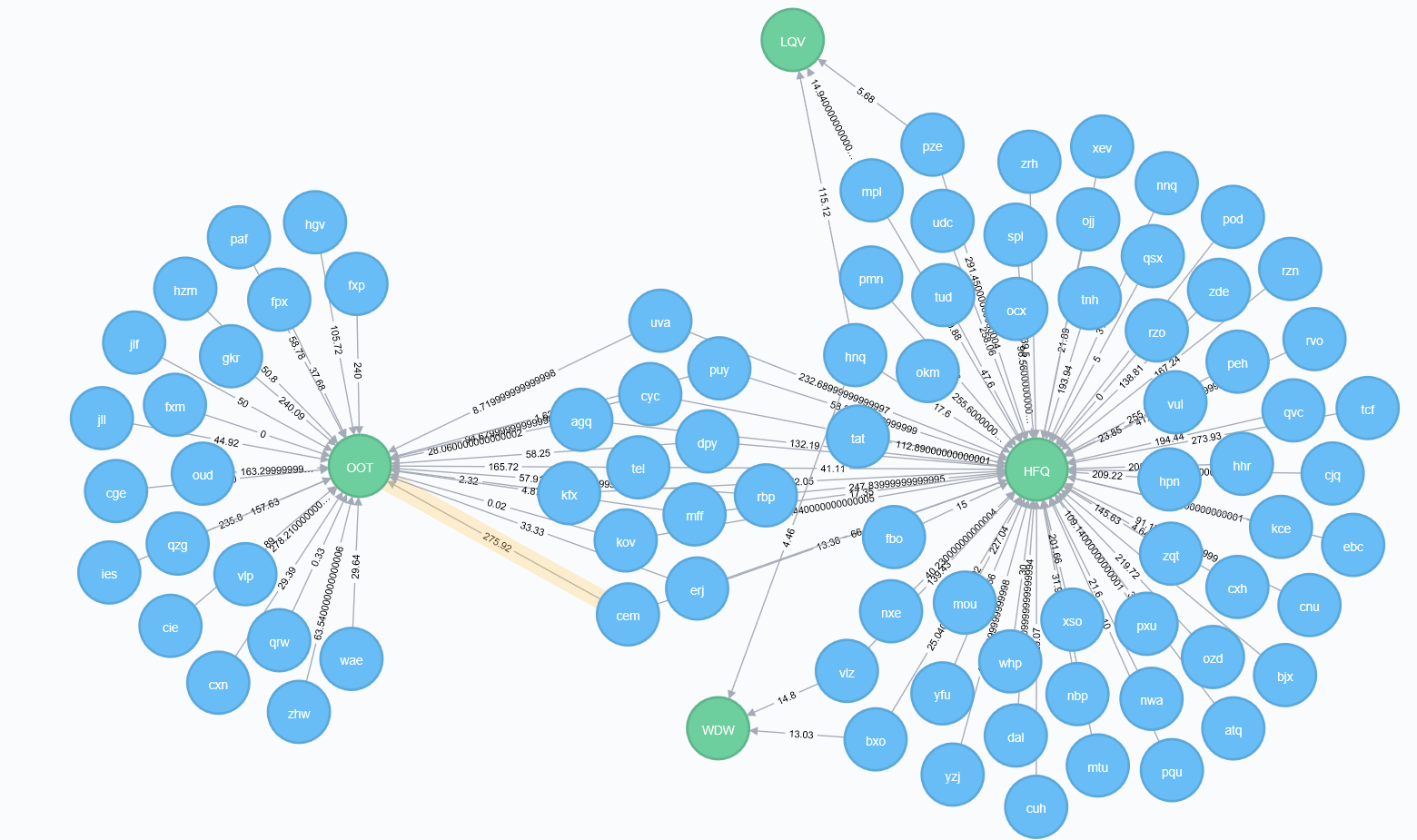


**Diameter: 13**

**Figure 8.** Visualization of developer network diameter

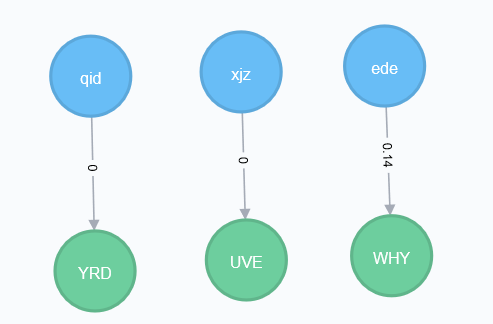
Closeness centrality, which uses the reciprocal of the average shortest distance to other nodes, is a measure of the degree to which an individual is near all other individuals in a network (Newman 2010). High closeness centrality therefore indicates close access to the resources in the network.

In the developer-project network, developer 'rbp', 'kfx', 'agq', 'tel', 'erj', 'kov', 'cyc', 'uva', 'dpy', 'mff', 'puy' and 'cem' are top twelve individuals having closest access. Furthermore, as it shown in figure 9, all of them work on project ‘OOT’ and ‘HFQ’. It can be therefore implied that project ‘OOT’ and ‘HFQ’ are projects gathering massive resource and by working in both, those twelve developers are able to get closer access to recourse.



**Figure 9.** Project ‘OOT’ and ‘HFQ’ are shared by twelve developers who have the highest closeness centrality.

In terms of nodes with smallest closeness centrality, as it shown in figure 10, developer ‘qid’, ‘xjz’, ‘ede’ all exclusively engage in their own individual project.

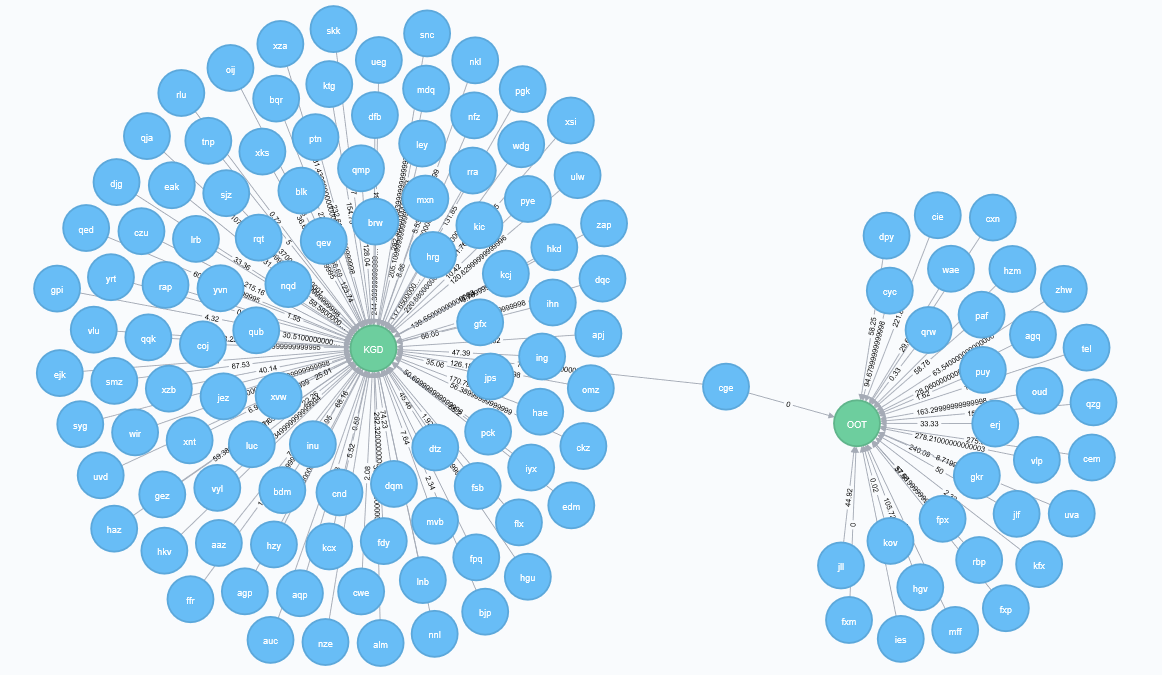


**Figure 10.** Developers with lowest closeness centrality and their contributions

* + 1. **Betweenness Centrality**

Betweenness centrality for each [node](https://en.wikipedia.org/wiki/Vertex_(graph_theory)) is the number of these shortest paths that pass through the node (Freeman 1977). Nodes with higher betweenness scores therefore have more control of other nodes in terms of communication access.

In the develop-developer network, developer ‘cge’ has the highest betweenness centrality. As shown in figure 11, ‘cge’ is the only person working in both ‘OOT’ and ‘KGD’, two principal projects in the company. Thus, it is reasonable to infer that developer ‘cge’ plays a significant role for the information sharing in between.

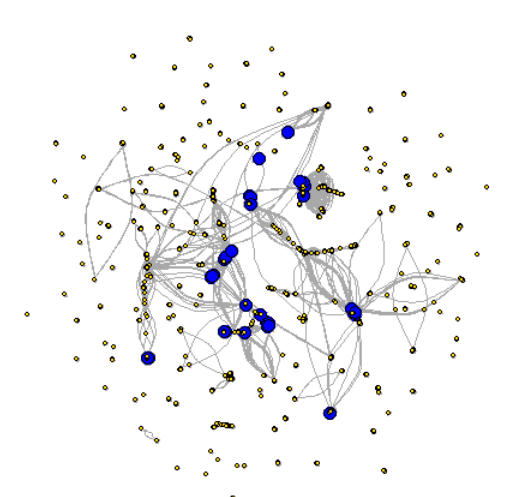


**Figure 11.** Developer ‘cge’ and corresponding connections

* + 1. **Eigenvector Centrality**

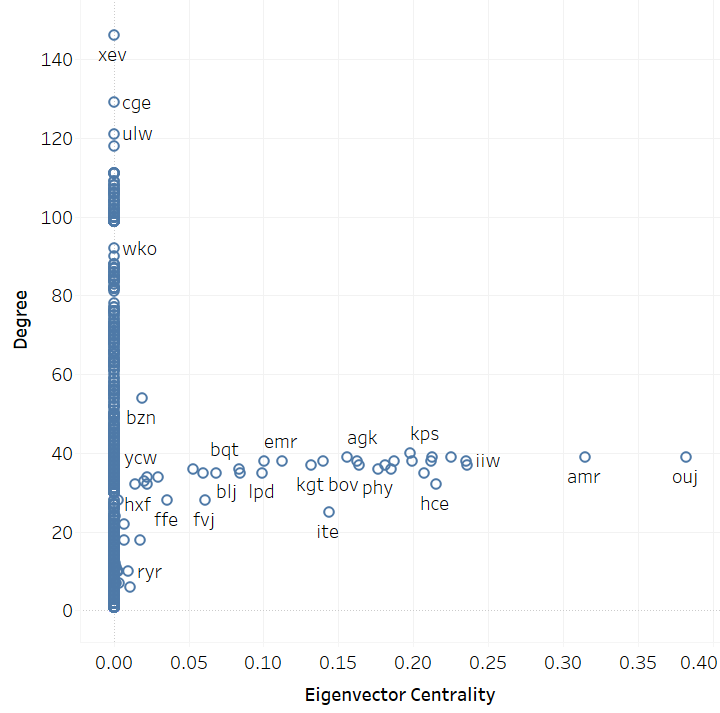
Eigenvector Centrality is a measurement revealing neighbors’ quantity and quality (Newman 2001). A high eigenvector score is resulted from a large number of high-quality neighbors. High-quality developers can be described as those involving in one or several principal projects, or participating in diversified projects, by which they interact with many people in the network. Correspondingly, low-quality developers are those engaging in unitary or individual projects, by which they collaborate with few others in the network.

As it shown in Figure 12, nodes colored in blue are those with highest Eigenvector. We find that they are widely positioned among different subcomponents in the graph and are connected with different sets of nodes.



**Figure 12.** Visualization of developers with high eigenvector scores

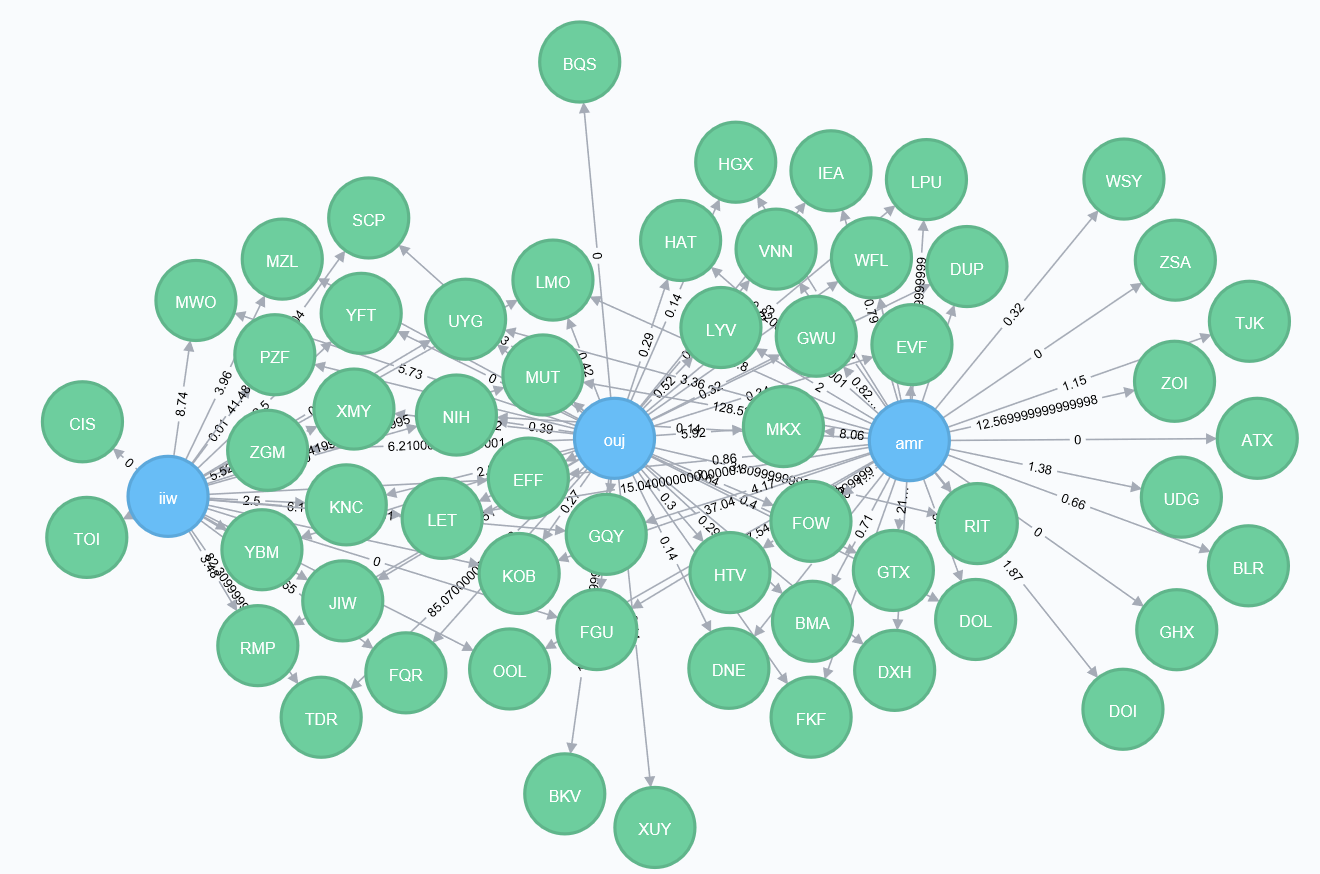
If looking at eigenvector and degree centrality together, as it shown in figure 13, we find that most nodes have an eigenvector centrality of about zero. Also, although that ‘xev’ has more than three times of degree than ‘ouj’, ‘ouj’ has the largest eigenvector centrality while ‘xev’ the least. In addition, other than ‘ouj’, developer ‘amr’ and ‘iiw’ are another two developers having high eigenvector centrality scores, while developer ‘xev, ‘cge’ and ‘ulw’ have the lowest.



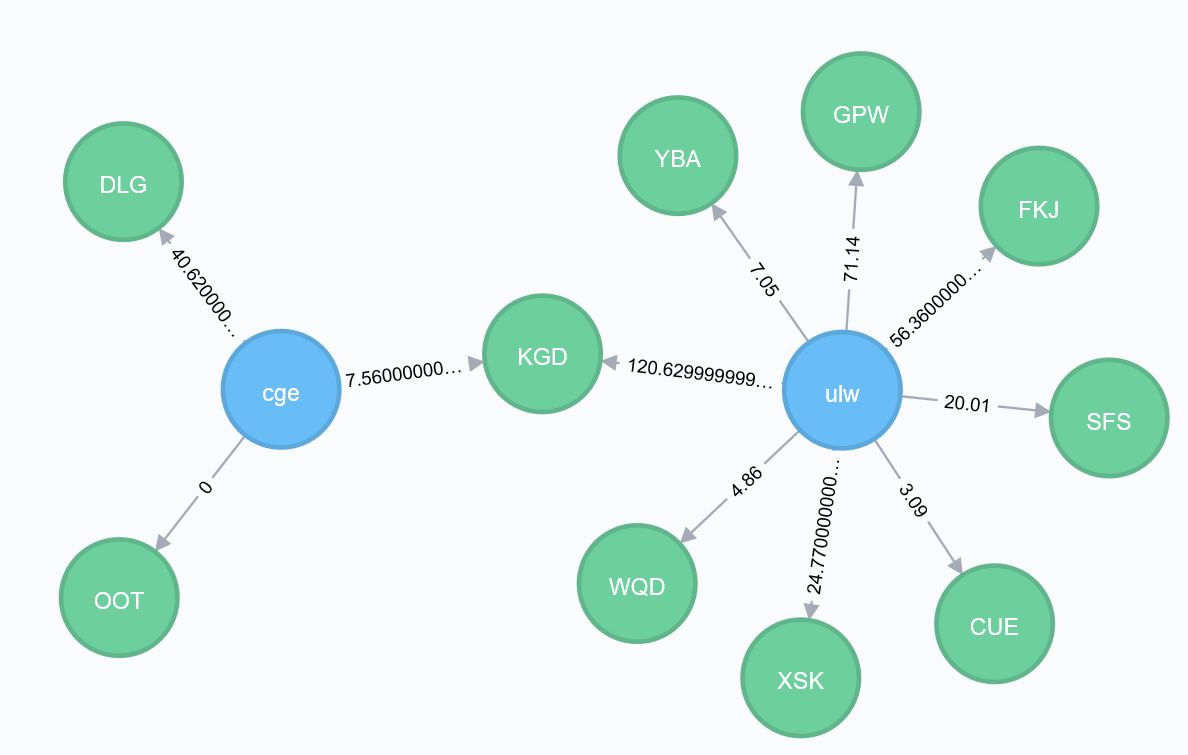
**Figure 13.** Scatter plot of degree and eigenvector centrality

Furthermore, as it shown in figure 14, developer ‘ouj’,’amr’ and ‘iiw’ intensively share projects. It can be therefore implied that those shared projects are worked by a lot of high-quality developers, bringing those three developers high-quality neighbors and thus high eigenvector centrality.

On the contrary, as it shown in figure 15, developer ‘cge’ and ‘ulw’ share only one project, ‘KGD’. It thus suggests that ‘KGD’ is a project worked by plenty of low-quality developers, giving both ‘cge’ and ‘ulw’ low-quality neighbors and thus low eigenvector centrality.



**Figure 14.** Developer ‘ouj’,’amr’ and ‘iiw’ share a lot of projects.



**Figure 15.** Developer ‘cge’ and ‘ulw’ share one large project ‘KGD’.

* 1. **Community Detection**

Communities in a network are groups of nodes internally connected or nodes sharing attributes (Girvan and Newman 2002). Detecting communities provides insights regarding the overall network structure, behavioral patterns of nodes and their relations (Fortunato 2010, Malliaros and Vazirgiannis 2013). For the developer-developer network, it supports findings about developers’ roles, collaboration preference as well as corporate cohesion. (Porter, Onnela and Mucha 2009)

However, community detection is an ill-defined problem. There is no universal definition of the methodology that one should follow. Consequently, there are no clear-cut guidelines on how to evaluate and compare the performance of different algorithms ([Fortunato](https://arxiv.org/find/physics/1/au:+Fortunato_S/0/1/0/all/0/1) and [Hric](https://arxiv.org/find/physics/1/au:+Hric_D/0/1/0/all/0/1) 2016).

Here we apply some popular and easy-accessible algorithms (Csárdi 2017; Peixoto 2014) and compare their performance. More details on network clustering can be found in several review articles (Chakraborty et al. 2016; Fortunato 2010; Malliaros and Vazirgiannis 2013; Satuluri et al. 2011; Porter et al. 2009; Xie et al. 2013).

Algorithms we apply can fall into three categories: optimization based, statistical inference based and dynamics based.

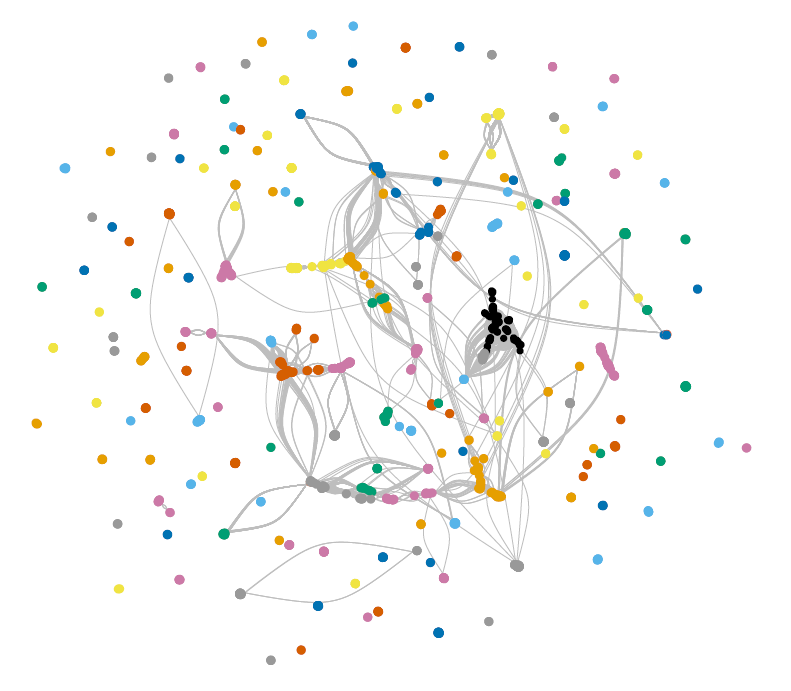
* + 1. **Optimization Based**

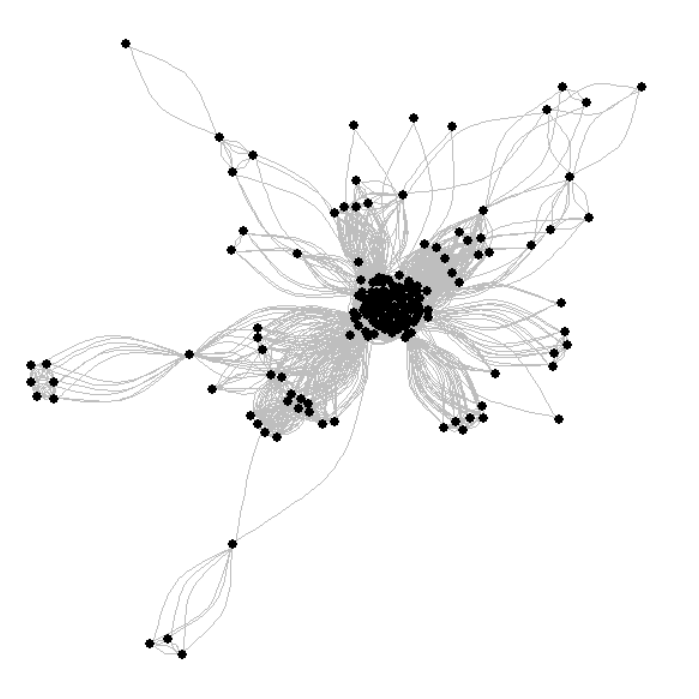
The idea of optimization based techniques is to optimize a quality function which measures the goodness of a clustering over all possibilities (Fortunato 2010). The most widely-used quality function is modularity by Newman and Girvan (Girvan and Newman 2006). It was designed to measure the strength of division of a network into communities. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

Based on the modularity measure, the multi-level modularity optimization algorithm is a hierarchical approach (Blondel et al 2008). It works according to the following steps: first, each node is assigned to a community independently. Afterwards, each node is moved to the community in a local, greedy way, where it achieves the highest contribution to modularity. When no node can be reassigned, each community is considered as a node on its own, and repeat the second step. The process stops either when there is only one node left or when the modularity cannot be increased any more.

The algorithm can be applied to weighted undirected graphs. The Multi-Level model generates a Modularity of 0.946. Groups (or communities) contain 12 developers on average. The largest community in the network, according to the model, contains 196 developers, and there are 73 ungrouped developers.

Figure 16 displays 223 groups with different colors, among which the largest group is colored in black.





**Figure 16.** Visualization of community detection in the developer’s network

The main problem with optimization based approach is that it ignores the nature of the considered community structures. In addition, there are 73 communities consists of only one developer. Suffering a resolution limit, modularity optimization methods are usually unable to detect small communities (Fortunato 2007).

* + 1. **Statistical Inference Based**

Unlike optimization based method, Statistical inference based method is a popular community detection way to generalize graph structures, by fitting [generative models](https://en.wikipedia.org/wiki/Generative_model) where parameters are inferred from data (Ball et al., 2011; Guimer`a and Sales-Pardo, 2009; Hastings, 2006; Karrer and Newman, 2011).

To find parameters for a best-fitted model, we use Minimum description length (MDL), which measures the amount of [information](https://en.wikipedia.org/wiki/Information_theory) required to describe the data. This approach corresponds to an implementation of [Occam’s razor](https://en.wikipedia.org/wiki/Occam%27s_razor), where the simplest model is selected, among all possibilities with the same explanatory power (Rissanen 1978).

The [stochastic block model](https://en.wikipedia.org/wiki/Stochastic_block_model) (SBM) is arguably the simplest generative process based on the notion of groups of nodes (Holland et al 1983). There are two types of [stochastic block model](https://en.wikipedia.org/wiki/Stochastic_block_model): non-degree-corrected and degree-corrected.

Non-degree-corrected model assumes that the edges are placed randomly inside each group, and as such the nodes that belong to the same group have very similar degrees. Therefore, the model usually performs bad on networks possessing highly heterogeneous degree distributions, while the degree-corrected model, on the other hand, provides a better fit for many empirical networks (Karrer and Newman 2011).

Table 6 shows the performance of two stochastic block models on the developer-developer network.

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Minimum description length** | **Group Number** |
| *Non-degree-correlated stochastic block model* | 68796.50778086044 | 52 |
| *Degree-correlated stochastic block model* | 67503.250630552488 | 61 |

**Table 6:** Degree correlated stochastic block model performs better than non-degree-correlated model based on minimum description length

A regular SBM has a drawback when applied to very large networks: it can’t find relatively small groups in very large networks. To address this problem, the nested stochastic block model (Peixoto 2014) was introduced. It not only finds small groups in large networks, but also provides a multilevel hierarchical description of the network. Table 7 demonstrates the general description of nested stochastic block models on the developer-developer network. Table 8 shows the group membership of developer ‘xon’ in different hierarchies.

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Minimum description length** | **Inferred Hierarchy** |
| *Non degree-correlated nested stochastic block model* | 58349.392019931613 | 7 |
| *Degree-correlated nested stochastic block model* | 58447.470967214511 | 6 |

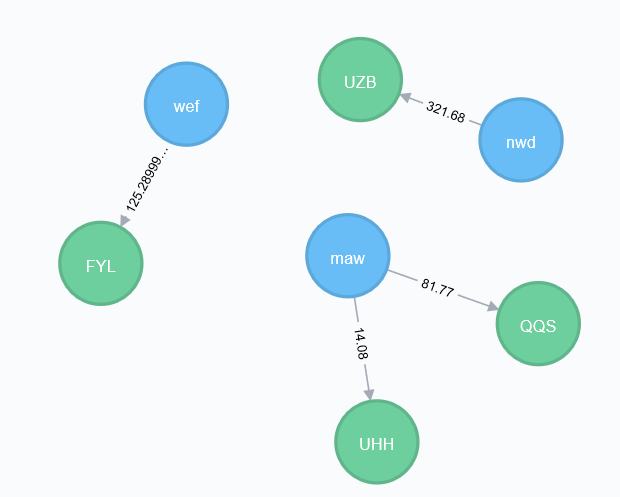
**Table 7:** Although it is often true that the degree-corrected model provides a better fit for many empirical networks, there are also exceptions. As it shown in table 7, non-degree-correlated nested stochastic block model generates lower minimum description length on the developer-developer network than the degree-correlated nested stochastic block model.

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Hierarchy** | **Group Membership** |
| *Non degree-correlated nested stochastic block model* | 0 | 14 |
| 1 | 8 |
| 2 | 7 |
| 3 | 6 |
| 4 | 5 |
| 5 | 4 |
| 6 | 3 |
| *Degree-correlated nested stochastic block model* | 0 | 13 |
| 1 | 0 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 0 |

**Table 8:** Developer ‘xon’ membership in different hierarchies

With the ability of generalizing graph structures, statistical inference based methods can be furthermore applied to predict missing or spurious edges, or potential edges in the network. The idea is generally that, based on the prior of observed graph, compute the posterior of graph with potentially connected edges, by sampling partitions from the posterior, then simulating the delete and insert of edges and eventually updating the likelihood (Clause 2008; Guimera 2005).

In the case of developer-developer network, detecting potential edges can lead to implications about which two developers are more likely to collaborate in the future. For example, there are no existing records about developer ‘maw’ collaborating with either ‘nwd’ or ‘wef’, as it shown in figure 17. With statistical inference model however, we can then get a probability for the collaboration (as it shown in table 9) and thus predict the likelihood of their future connection.



**Figure 17.** Developer ‘maw’ has no collaboration with either ‘nwd’ or ‘wef’.

|  |  |  |
| --- | --- | --- |
| **Algorithm** | **Potential collaboration** | **Probability** |
| *Degree-correlated nested stochastic block model* | ‘maw’ and ‘ nwd’ | 55.55% |
| ‘maw’ and ‘wef’ | 44.45% |

**Table 9:** ‘maw’ is slightly more likely to collaborate with ‘nwd’, rather than ‘wef’, based on the degree-correlated nested stochastic model.

Here we apply degree-correlated nested stochastic block model to demonstrate general nonparametric statistical models’ ability to infer missing or potential edges. More details on network completion problem can be found in several review articles (Rodriguez et al 2010; Guimer´a et al 2005; Hanneke et al 2009).

* + 1. **Dynamics Based**

Other than statistical and optimization, communities can also be identified by methods of running dynamical processes on the network (Fortunato 2010), among which Random walk dynamics is by far the most exploited.

Walktrap and Infomap are two algorithms both applying random walk, with the former based on an observation that short random walks have the tendency of staying in the same community, while the latter aimed at minimizing the expected description length of a random walk trajectory (Pons and Latapy 2006).

Table 10 shows the comparison of the model performance between Walktrap and Infomap.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Algorithm** | **Modularity** | **Group Number** | **Average Group size** | **Largest Group size** | **Ungrouped Developer Number** |
| Walktrap | 0.925 | 277 | 9.46 | 152 | 73 |
| Infomap | 0.901 | 303 | 8.65 | 77 | 73 |

**Table 10:** Comparison of Walktrap and Infomap

Infomap and walktrap model have the same ungrouped developer number, as well as similar modularity and average group size. However, the largest community in walktrap model contains 152 developers, twice of the counterpart in Infomap model.

A comprehensive and optimized way of inferring network structure through community detection is beyond the scope of this paper. Here we apply some easy-accessible algorithms, namely multi-level model, [stochastic block model](https://en.wikipedia.org/wiki/Stochastic_block_model)s, walktrap and Infomap, which can be categorized as optimization based, statistical inference based and dynamics based methods to the developer-developer network. When comparing model performance, we turn to modularity, minimum description length, as well as group number, average group size and ungrouped node number for reference.

## Summary

To understand the interaction among developers, we create a weighted developer-developer graph by projecting the weighted bipartite developer-project graph. Four graph properties, namely degree, closeness, betweenness and eigenvector centrality, are studied, from which we achieve implications about developers’ functionality, roles and relationships. In addition, we compare and implement different community detection algorithms, such as popular modularity maximization approach, to categorize and capture the network structure.

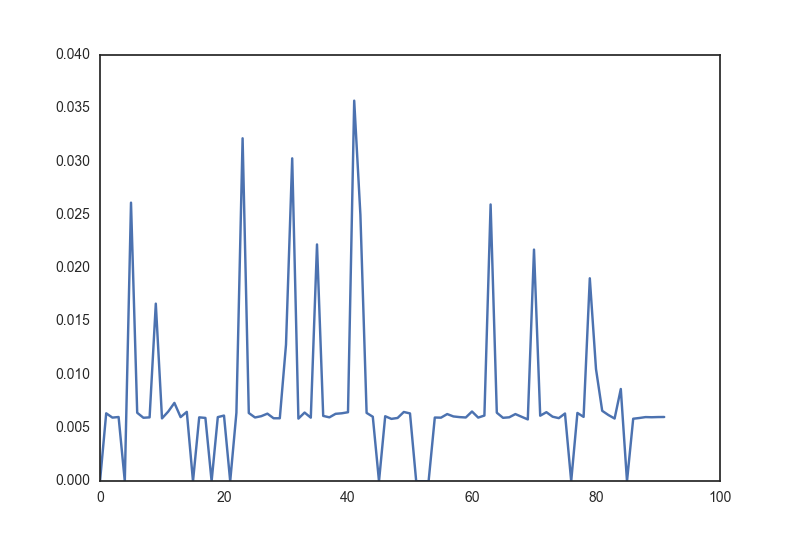
# Modeling Dynamics: Evolving Network

Analyzing dynamic network has gained increase popularity and significance as longitudinal network data becomes more available. Organizations study corporate network in a dynamic way in order to discover the decision cycle of major events, track the evolutions of corporate finances, operations and culture, as well as detect changes in the organizational behavior. Being able to identify a change in network dynamics can enable managers to better prepare for the coming change and prevent potential problems. (Doreian and Stokman 1997)

Dynamic network analysis (DNA) brings together traditional social network analysis, link analysis, social simulation and multi-agent systems within network science and network theory.

(Kathleen 2014). There are two aspects of this field. The first is the statistical analysis of DNA data. The second is the utilization of simulation to address issues of network dynamics. Dominant modeling methods include Markov chain models, multi-agent simulation models, and statistical models. (McCulloh and Carley 2009)

A complete application of dynamic network modeling is beyond the scope of this paper. Considering the nature of the given problem, we choose statistical models as our major measure. In other words, we compare the properties of networks at different points in time.

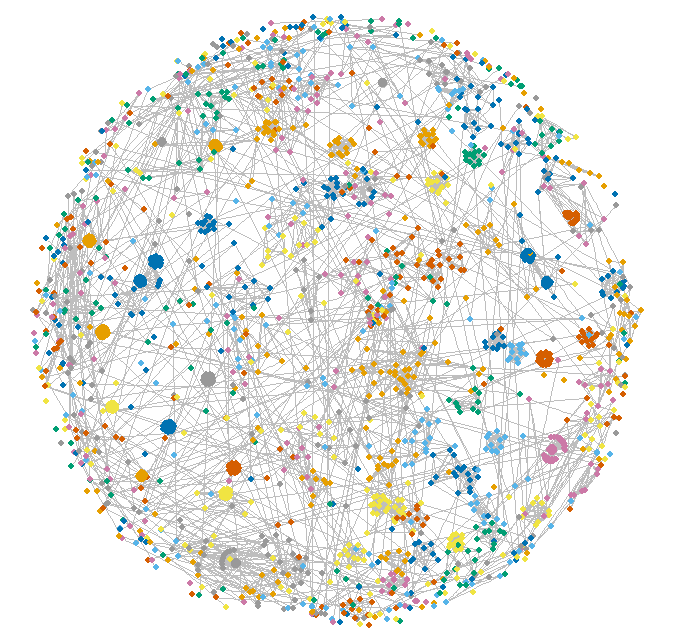
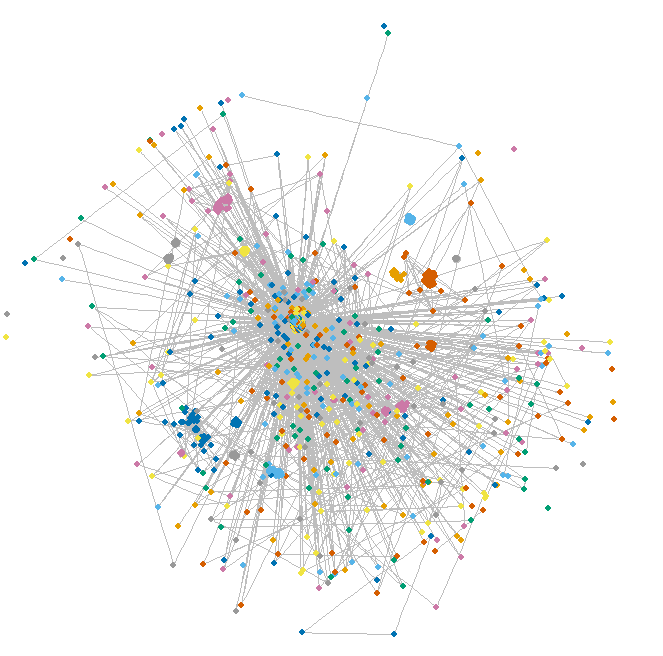


**Visualize days with density zero and days with high density.**

**Compare community numbers, clustering coefficient**

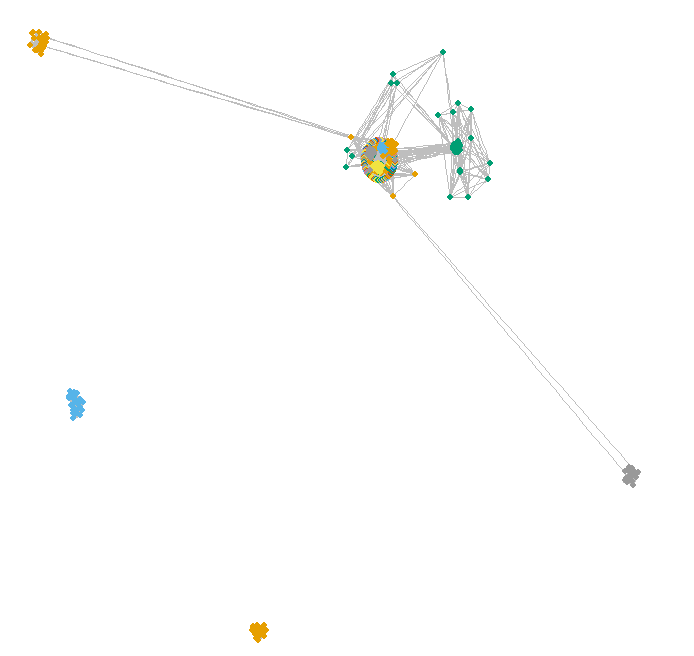
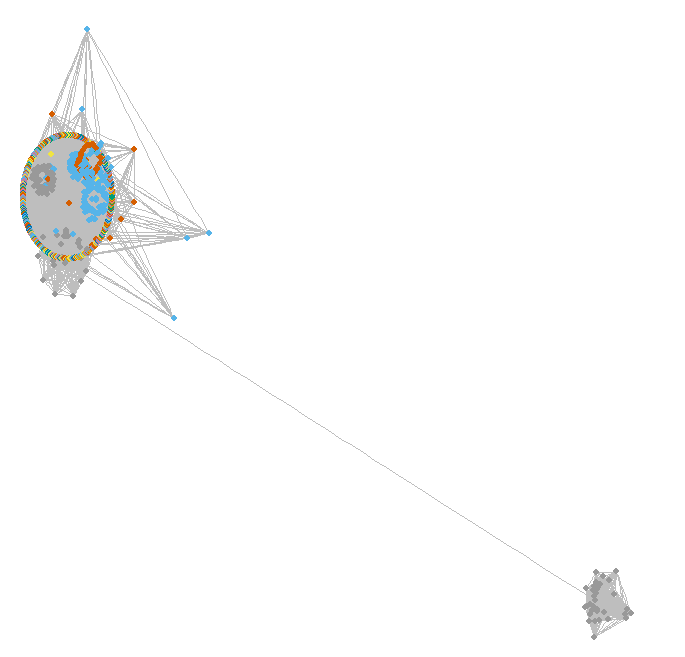
**Day 1**

**Day 0**



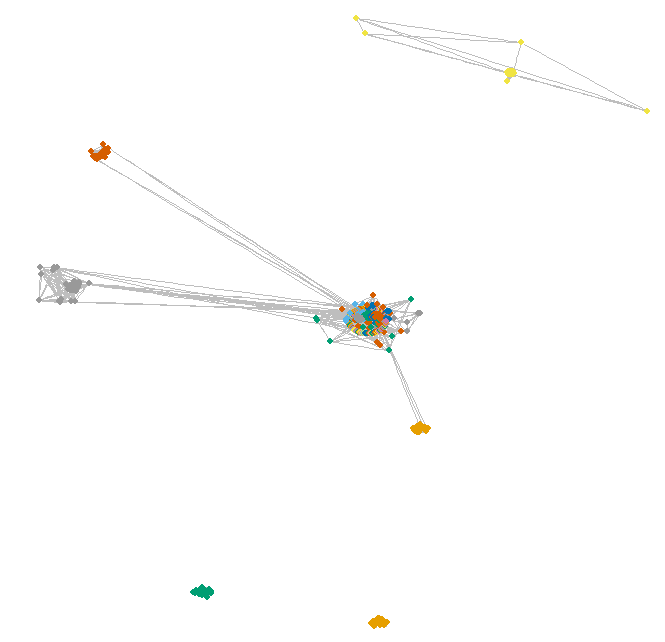
**Day 3**

**Day 2**



**Day 5**

**Day 4**



# Discussion and concluding remarks

Network data science is the use of data science methods, tools and algorithms in the modeling and analysis of network (graph) data. We demonstrated how a network data science approach can help companies gain rich insights about employee effort, contribution, and collaboration. These insights are useful in order to optimize work patterns and productivity in organizations.

# A study of the proprietary software development was started with constructing weighted graphs. Afterwards, initial insights about developers’ role (managers or engineers), productivity and individual influence were achieved through analyzing network properties. Finally, implications about software development cohesion and culture were developed through implementation of cluster analysis as well as a comparison between constructed networks and complex systems.

In addition to the analysis of network data, we proposed a recommender system that managers could consult to assign employees to projects…

# In addition to the analysis of a snapshot or aggregated version of the software development network at this particular company, we did a longitudinal and dynamic analysis. This provides better understanding of how node attach and detach from the network (developers dropping off or initiating projects), and thus lead to a more thorough understanding of developers’ working pattern.

There are many opportunities for future research to extend the methodology proposed in this paper. …

# References

Albert R., Barabasi A.-L. 2002. Statistical mechanics of complex networks. Rev. Mod. Phys. 74: 47–97.

Ball, B., Karrer B., and Newman M. E. J. 2011. An efficient and principled method for detecting communities in networks. Phys. Rev. E 84, 036103.

Blondel Vincent D., Guillaume Jean-Loup, Lambiotte R., Lefebvre E. 2008. Fast unfolding of communities in large networks. [arXiv:0803.0476](https://arxiv.org/abs/0803.0476) [physics.soc-ph].

Breese J.S., Heckerman D., and Kadie C. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In Proceedings of the Fourteenth Conference on Uncertainty in Artifical Intelligence.

Brian Karrer, and Newman M. E. J. 2011. Stochastic Blockmodels and Community Structure in Networks. Physical Review E 83, no. 1 (2011): 016107

Chakraborty T., Kumar S., Ganguly N., Mukherjee A., Bhowmick S. 2016, GenPerm: A Unified Method for Detecting Non-overlapping and Overlapping Communities. preprint arXiv:1604.03512.

Clauset, A., Shalizi, C. R., Newman, M. E. J. 2009. Power-Law Distributions in Empirical Data. SIAM Review 51(4):661-703, 2009.

Clauset A., Moore C., Newman M. E. J. 2008. Hierarchical structure and the prediction of missing links in networks. Nature 453, 98-101 (2008), DOI: 10.1038/nature06830

Csárdi G. et al 2017. Network Analysis and Visualization. http://igraph.org

Das A., Datar M., Garg A., and Rajaram S. 2007. Google News Personalization: Scalable Online Collaborative Filtering. International World Wide Web Conference, Proceedings of the 16th international conference on World Wide Web

Doreian, P., and Stokman, F.N. (Eds.) 1997. Evolution of Social Networks. Amsterdam: Gordon and Breach.

Fan Y., Li M., Zhang P., Wu J., Di Z. 2007. The effect of weight on community structure of networks. PHYSICA A 378 (2007) 583–590

Fortunato, S., Barthelemy, M. 2007. Resolution Limit in Community Detection. PNAS 104, 36- 41.

Fortunato S. 2010. Community detection in graphs. Phys. Rep. 486 (3–5): 75–174.

[Fortunato](https://arxiv.org/find/physics/1/au:+Fortunato_S/0/1/0/all/0/1) S., [Hric](https://arxiv.org/find/physics/1/au:+Hric_D/0/1/0/all/0/1) D. 2016. Community detection in networks: A user guide. arXiv:1608.00163 [physics.soc-ph]

Freeman, Linton 1977. A set of measures of centrality based on betweenness. Sociometry. 40: 35–41.

Girvan M., Newman M. E. J. 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA. 99 (12): 7821–7826.

Guillaume, J., Latapy, M. 2004. Bipartite structure of all complex networks. Information Processing Letters, 90(5):215-221.

Guimer`a, R., and Amaral L. A. N. 2005. Missing and spurious interactions and the reconstruction of complex networks. Nature 433, 895.

Hanneke S., Xing E. 2009. Network completion and survey sampling. In AISTATS ’09.

Hastings, M. B. 2006. Community Detection as an Inference Problem. Phys. Rev. E 74(3), 035102.

Hazewinkel, Michiel, ed. 2001, Kolmogorov–Smirnov test,  [Encyclopedia of Mathematics](https://en.wikipedia.org/wiki/Encyclopedia_of_Mathematics), Springer, [ISBN](https://en.wikipedia.org/wiki/International_Standard_Book_Number) [978-1-55608-010-4](https://en.wikipedia.org/wiki/Special:BookSources/978-1-55608-010-4).

Holland Paul W., Laskey Kathryn B., Leinhardt S. 1983. Stochastic blockmodels: First steps, Carnegie-Mellon University, Pittsburgh, PA 15213, U.S.A.

Kathleen M. Carley. 2014. ORA: A Toolkit for Dynamic Network Analysis and Visualization. Encyclopedia of Social Network Analysis and Mining, Springer.

Malliaros F. D., Vazirgiannis M. 2013. Clustering and community detection in directed networks: A surve. Phys. Rep. 533 (4): 95–142.

Marsaglia G, Tsang WW, Wang J. 2003. Evaluating Kolmogorov's Distribution. Journal of Statistical Software. 8 (18): 1–4.

McCulloh Ian, Carley K. 2009. Longitudinal Dynamic Network Analysis. CMU-ISR-09-118

[Newman M. E. J. 2001. Scientiﬁc collaboration networks. II. Shortest paths, weighted networks, and centrality. PHYSICAL REVIEW E, vol. 64, 016132.](http://www-personal.umich.edu/~mejn/papers/016132.pdf)

Newman, M.E.J. 2010. Networks: An Introduction. Oxford, UK: Oxford University Press.

Newswire 2013. BlueOptima Coding Effort Analytics Announces Support for Git. Retrieved from <https://www.newswire.com/blueoptima-coding-effort-analytics/252989>.

Peixoto Tiago P. 2014. The graph-tool python library. doi:10.6084/m9.figshare.1164194 Key: citeulike:13701202.

Peixoto Tiago P. 2014. Hierarchical block structures and high-resolution model selection in large networks. Phys. Rev. X 4, 011047.

Pons, P., Latapy M. 2006, Computing Communities in Large Networks Using Random Walks. International Symposium on Computer and Information Sciences (Springer), pp. 284– 293.

Porter M. A., Onnela J.-P., Mucha P. J. 2009. Communities in Networks, Math. Soc. 56: 1082–1097, 1164–1166.

Ricci F., Rokach L., Shapira b. 2011. [Introduction to Recommender Systems Handbook](http://www.inf.unibz.it/~ricci/papers/intro-rec-sys-handbook.pdf), Recommender Systems Handbook, Springer, 2011, pp. 1-35

Rissanen, J. 1978. Modeling by shortest data description. Automatica. 14 (5): 465–658.

Rodriguez M. Gomez, Leskovec J., Krause A. 2010. Inferring networks of diffusion and influence. KDD ’10.

Satuluri V., Parthasarathy S. and Ruan, Y. 2011 Local Graph Sparsification for Scalable Clustering. In: Proceedings of the 2011 International Conference on Management of Data, ACM Press, New York, 721-732.

Xie J., Kelley S., and Szymanski B. K., 2013, Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput. Surv. 45(4), 43:1.

Zhan, Benjamin F., Noon, Charles E. 1998. Shortest Path Algorithms: An Evaluation Using Real Road Networks. Transportation Science. 32 (1): 65–73. doi:10.1287/trsc.32.1.65.

Zhou T., Ren J., Matus M., Zhang Yi-Cheng. 2007. Bipartite network projection and personal recommendation. PHYSICAL REVIEW E 76(4): 046115.