**Validating a new flocking based evolutionary computation strategy for measuring centrality of online social networks**

Student: Christian Johnston, Supervisor: Professor Iain A. Stewart

Submitted as part of the degree of MSci Natural Sciences to the Board of Examiners in the School of Engineering and Computing Sciences, Durham University

***Abstract —***

**Context/Background**

Online social network (OSN) sites like Facebook and Twitter are among the most popular sites on the internet which provides an opportunity to study the characteristics of their networks at a large scale through social network analysis. Understanding these networks is important, both to improve current systems and to design new applications of OSNs. Centrality is one of the main research topics in social network analysis and there are a number of traditional methods of measuring centrality, each with their own uses and limitations.

**Aims**

The project aims to compare 4 traditional methods of measuring centrality, illustrating their uses but also their limitations when applied to online social networks highlighting the need for a new measure with characteristics more realistic and practical in OSNs. The projects main aim is to validate a flocking based technique proposed by Dhinesh Babu L.D and Ebin Deni Raj in 2015 both theoretically and empirically through the use of consistent data-sets to determine whether it provides an accurate and more realistic measure of centrality when applied to OSNs.

**Method**

The new flocking based method will be developed and implemented using Boids program created by Craig Reynolds in 1989 and will be applied to random artificial networks of various sizes as well as to benchmark data-sets used in the original paper. The new method, which will be referred to as flocking based centrality for social networks (FBCS) algorithm, will be compared to the traditional measures using Kendall’s tau correlation and robustness.

**Proposed Solution**

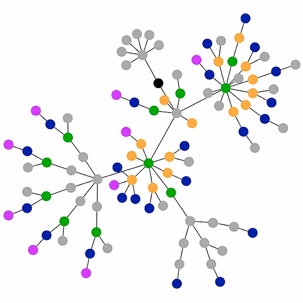
After interpreting both the pseudo-code and narrative in the paper of Babu and Raj so as to derive less ambiguous algorithms, a full and precise implementation of the subsequent algorithm will be developed, along with implementations of other more standard algorithms. The artificial data-sets produced will closely mimic the characteristics of real OSNs providing sufficient data for social network analysis through tools such as NetworkX and Gephi. Numerous analytical experiments will be conducted including the varying of algorithm parameters in an attempt to achieve more accurate results.

***Keywords —***Network Theory, Graph theory, Social Network Analysis, Boids algorithm, Evolutionary Computation, Centrality, Connected Graphs.

**I**  **INTRODUCTION**

**A Online Social Networks**

Online social networks (OSNs) are defined as web-based services that allow individuals to (1) construct a public profile within a system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. (Danah M. Boyd, 2007)



***Figure 1****: Image representing an OSN exhibiting particular characteristics such as the formation of clusters.* (See section II.D for further details)

Image sourced from: *http://authenticorganizations.com/harquail/2011/01/24/your-authentic-social-network-the-identity-graph/*

In an OSN, the spread of information is affected by the influence of certain nodes. Research results show that these influential nodes are rare in social networks, but their influence can quickly spread to most nodes in the network and as such, these nodes greatly influence the structure and function of networks. Identifying these influential nodes helps to better understand complex networks and can help in accurately predicting and controlling network evolution. (Xiaohui Zhao, 2017)

One of the key features of an OSN is that its growth is random and extremely dynamic, relating to the fact that someone can enter a new friend circle or make a new connection unrelated to their past. Furthermore, all OSNs have scale-free[[1]](#footnote-1) properties and there is no hierarchy between nodes and so any node can connect to any other.

**B Social Network Analysis**

Social network analysis (SNA) is a branch of network analysis and graph theory that examines the structure of relationships between social entities. It characterises networked structures in terms of nodes (people, organizations, groups or websites) and the ties, edges, or links that connect them. These networks are often visualized through ‘sociograms’ where nodes are represented as points and ties are represented as lines.

SNA has become increasingly important in the last decade due to the massive increase in the number of social network users and therefore the rapidly increasing amounts of data available to be analysed. Each second, billions of bytes of data are generated in social networks and this huge quantity of information can be analysed to provide insight into research domains such as news spreading, friendship networks, collaborative rating, disease transmission, page rank and centrality.

**C Centrality**

In graph theory and network analysis, centrality is the comparative importance of nodes in a network graph. Relating specifically to SNA, centrality measures the extent to which an individual interacts with other individuals in the network; the more an individual connects to others in a network, the greater their centrality. Applications of this are identifying the most influential person in an OSN and for discovering the nodes which can influence the maximum number of users.

It is necessary to differentiate between ‘local’ and ‘global’ point centrality. A point is locally central if it has a large number of connections with the other points in its immediate environment and a point is globally central when it has a position of strategic significance in the overall structure of the network (Scott, 1991). There are many traditional methods (*see section II E*) to measure centrality of a network node, each attempting to capture some set of intuitions and as a result, each has its own uses and limitations. (Disney, 2014)

**D Project Purpose, Aims and Research question**

The purpose of the project in summary is to validate a new measure of centrality proposed by Dhinesh Babu L.D and Ebin Deni Raj (2015) both theoretically and empirically. The project will aim to compare the statistical distribution of the new centrality measure with respect to traditional centrality measures using artificial networks of varying sizes and benchmark data-sets.

The deliverable should be an executable program that is capable of analysing any connected network; calculating the centrality of the graph’s nodes using the new measure. The algorithm should be deterministic and ideally will be efficient such that it is capable of analysing networks with up to 100,000 nodes in a reasonable time and have computational complexity less than O().

**Research Question**: Does the new centrality measure proposed by Babu and Raj in 2015, based on Boids algorithm, provide an accurate and more realistic measure of centrality when applied to online social networks in comparison with traditional measures?

**E Deliverables (***unfamiliar concepts introduced here will be discussed later)*

**Minimum Objectives:**

* Create a basic network of nodes representing a social network for proof of concept of the traditional centrality measures.

Choose appropriate social network analysis software; both tool built for scripting language as one based on a graphical user interface (GUI).

* Evaluate qualitatively the limitations of traditional centrality methods highlighting the need for a new method with parameters more suited to OSNs.
* Download all benchmark datasets from the internet in required format.
* Implement a basic program that will take as an input a network data-set and will calculate the node of highest centrality using the traditional measures utilising functions provided by my chosen SNA software package.

**Intermediate Objectives:**

* Write procedures to automate the process of calculating interaction, separation and cohesion: the 3 rules used in Boids to implement the FBCS algorithm.
* Apply the flocking based centrality method as well as the traditional measures to the benchmark data-sets used in the paper for comparison.
* Evaluate the FBCS method against traditional methods using statistical correlation coefficients.
* Evaluate the traditional centrality measures and FBCS algorithm in terms of computational complexity.
* Compare obtained results against results in original paper and analyse effectiveness of implementation.

**Advanced Objectives:**

* Create artificial networks of given sizes using a known random network model and apply the FBCS algorithm along with the traditional measures of centrality to these data-sets.
* Explore how the topological structure of networks and dissimilarity between connected nodes neighborhoods impacts a nodes centrality and how the flocking based algorithm could be altered to account for this with the inclusion of a new procedure.

**II DESIGN**

**A Description of Software Development Life Cycle**

Due to the nature of the chosen architectural style (*see section G*), it is essential that all of the three procedures are implemented before the algorithm is applied to data-sets. For this reason, the project is more suited to a linear, sequential process model such as the Waterfall model allowing me to fulfil the functional requirements identified in the analysis stage of the software development life cycle. This approach has benefits of easy management, enforced discipline on meeting requirements and provides an easy measure of progress.

However, once the main algorithm is developed, due to the modularity of the architectural style, I plan to alter and add parameters in order to obtain better results in the ranking of the nodes of a given network and to try and achieve improved performance. As a result, I will then turn to an iterative, agile development methodology when optimising the algorithm, seeking to continually improve upon previous attempts. As a result, the project will benefit from a reduced product cycle time as well as allowing the continual testing leading to incremental improvements enabling me to fulfil the non-functional requirements which are more focused on performance and the extension of the algorithm.

**B Requirements**

Due to the algorithmic nature of the program, there are a limited number of functional requirements (FR). Thus, there is a focus on non-functional requirements (NFR) relating to the performance of the algorithm both in terms of efficiency and the values obtained in comparison to the original paper.

***Table 1****: Functional & Non-Functional Requirements*

|  |  |  |
| --- | --- | --- |
| **ID** | **Requirement** | **Priority** |
| FR1 | Application of traditional centrality measures to benchmark data-sets. | High |
| FR2 | Visualisations of network graphs using Python’s built in Matplotlib library, ranking each nodes size by its centrality value for a given measure. | High |
| FR3 | Utilise the advanced visualisation features provided by the chosen GUI package. | Medium |
| FR4 | Implementation of my interpretation of the FBCS algorithm utilising the three rules in Boids as a framework for the three procedures, following the chosen architectural style. | High |
| FR5 | Create artificial networks of varying sizes using known random network models. | Medium |
| NFR1 | Optimisation of algorithm by tweaking parameters used. | Medium |
| NFR2 | Explore how the topological structure of networks and dissimilarity impacts a nodes centrality. | Medium |
| NFR3 | Evaluate all centrality measures in terms of computational complexity. | High |
| NFR4 | Evaluate the correlation between traditional measures and the FBCS measure using a statistical correlation coefficient. | High |
| NFR5 | Utilise existing knowledge from associated published literature to review the role of each centrality measure in the determination of influential nodes. | Low |

**C Specification of Software**

Social network analysis software facilitates quantitative and qualitative analysis of social networks by describing features of a network either through numerical or visual representation and is either based on graphical user interfaces or built for scripting languages. As I plan to develop my own centrality measure whilst still being able to utilise the powerful features of the SNA software, I plan to use packages built for scripting languages due to their extensibility rather than those based around a GUI.

Commonly used scripting tools used for network analysis include: NetMiner, igraph, MuxViz (based on R) and the NetworkX Python library. The chosen scripting based software for the project will be the NetworkX Python library which I have selected for multiple reasons.

Firstly, Python is a powerful programming language that allows simple and flexible representations of networks as well as concise expressions of network algorithms. Other benefits of Python include rapid prototyping, clear code style and its platform-neutrality meaning the final implementation will run on almost any platform without the need to recompile the source. This makes it easy to continue development from a variety of computers during the project, especially in the event that additional processing power is preferred later on to run on large data-sets. Secondly, Python is chosen because of its familiarity, enabling a reduced implementation period in the system development life-cycle, allowing more time to be spent on the design, planning and evaluation sections of the cycle.

Due to its dependence on a pure-Python "dictionary of dictionary" data-structure, NetworkX is an efficient, scalable and portable framework for SNA, allowing operation on large real-world graphs in excess of 10 million nodes and 100 million edges. NetworkX has built in visualisation tools via the Matplotlib library however this is limited and as a result, a GUI package may be better suited such as NetMiner, UCINet, Pajek or Gephi. I have decided to go with Gephi as it is the most popular network visualisation package on the market, it doesn't require any programming knowledge and it is open-source and free.

**D Data Sets**

To evaluate the effectiveness and the efficiency of the proposed centrality measure, I will apply it to both real and artificial networks. The data-sets used in the original paper which I will use for comparison are shown in Table 2.

***Table 2****: Benchmark data-sets used for comparison ranked by size.*

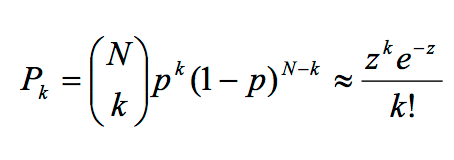
|  |  |  |
| --- | --- | --- |
| **Name of data-set** | **Number of Nodes** | **Number of edges** |
| Zachery Karate club network (Zachary, 1977) | 34 | 78 |
| Dolphins Social network (D. Lusseau, 2003) | 62 | 159 |
| American Football Network (Girvan, 2002) | 115 | 613 |
| Celegan’s Neural Network (Hilgetag, 2006) | 297 | 1359 |
| Political Blog Network (Glance, 2005) | 1490 | 19090 |

One consideration is that ‘Celegan’s Neural Network’ and the ‘Political Blog Network’ are directed graphs however many of the built-in NetworkX functions only work on undirected graphs and so I have made the assumption that all networks are undirected in my abstraction of the problem. As a result, conversion from directed to undirected is essential for the above network data-sets. I have followed the convention that if node has a directed edge to node and node has a directed edge to node , then the undirected graph will have one undirected edge between the two nodes as opposed to two-undirected edges, which is the convention used in the original paper.

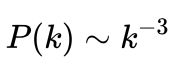
These benchmark data-sets are small OSNs and so to fully test the new centrality measure it will be necessary to create artificial networks of chosen size through the use of random network models. It is essential that these artificial networks exhibit the properties one would find in real-world social networks as the new centrality measure is specifically tailored for analysing OSNs.

There are three distinctive features of network structure in OSNs. The first of these is the ‘small-world’ effect; many pairs of apparently distant people are actually connected by a short chain of intermediate acquaintances. In relation to this, the *‘6 Degrees of Separation’* was proposed by Frigyes Karinthy in the early 20th century. Secondly, all social networks exhibit clustering; the probability of a tie between two actors is much greater if the two actors in question have another mutual acquaintance. This is defined by a ‘clustering coefficient’, which is the probability that two acquaintances of a randomly chosen person are themselves acquainted. Thirdly, all social networks exhibit a skewed degree distribution. (M. E. J. Newman, 2002)

In many fields, the Erdös-Renyi (ER) model[[2]](#footnote-2) is considered as the basic model in the study of random networks, however, networks generated by this model have a low clustering coefficient and a Poisson degree distribution, rather than a skewed power law as observed in many real-world OSNs (Strogatz, 1998). The Watts-Strogatz (WS) model addresses the first of these two limitations; it accounts for clustering whilst retaining the short path lengths of the ER model. Consequently, the model is also able to explain the "small-world" phenomena. The major limitation of the model is that it produces an unrealistic degree distribution as shown in Figure 2 (Strogatz, 1998). Such networks are better described in that respect by the Barabási–Albert (BA) model as shown in Figure 3. On the other hand, the BA model fails to produce the high levels of clustering seen in real networks, a shortcoming not shared by the WS model (Barabási, 2002). Thus, neither the WS, nor the BA model should be viewed as fully realistic.

***Figure 2****: Poisson degree distribution generated by Watts-Strogatz model. Where N = Number of nodes, p = probability that two nodes are connected, k = degree of a node.*

***Figure 3****: Scale-free degree distribution generated by the Barabási–Albert model where k = degree of a node.*

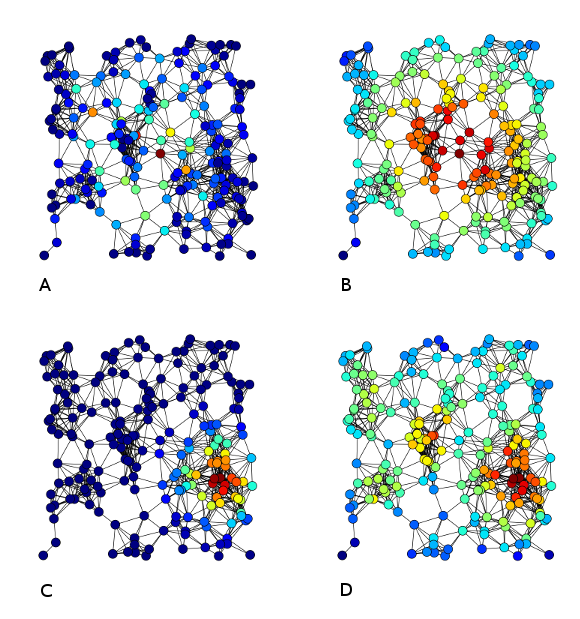


NetworkX helps abstract the detail of these models by providing functions which create random networks of chosen size based on all three of these models. After studying numerous related literature, I have decided that the Watts-Strogatz model is the most suitable for constructing small-world networks and so this is the model I will be using.

When analysing and visualising networks of up to a million nodes, a large amount of computing power is required for the program to finish execution in a reasonable amount of time as many of the traditional centrality measures are of computational complexity O() where is the number of nodes. As a result, I will look to utilise the CIS High Performance Computing service called Hamilton provided by Durham University. This is a Linux cluster based on Intel processors and IBM GPFS parallel file systems and is dedicated to solving numerically intensive programming problems.

**E Traditional Centrality Measures**

In 2005, Borgatti stated that centrality index will be different for various types of networks and that each index will be appropriate for a particular instance (Borgatti, 2005). Moreover, each of the traditional measures have specific uses as each measure is attempting to capture some set of intuitions so each will be of limited use to address questions where some other intuition is needed. We will look at 4 of the most common measures of centrality.

Degree centrality **(D)** is based on the number of direct connections a node has with other nodes and thus it often known as the local topological property of a node in the network. As such, this measure of centrality can only be used to show how well-connected points are within their local environments, not how central they are globally. (Scott, 1991)

Betweeness centrality **(A)** measures the extent to which a node lies ‘between’ the various other nodes in the network. It is a measure of the degree to which a node serves as a bridge by measuring the number of times a node lies on the shortest path between other nodes. As a result, it is often used for community detection in complex networks.

Closeness centrality **(B)** is a measure of the shortest distance between a given node and all other nodes. In a flow context, closeness can be interpreted as the minimal time until arrival of something flowing through a network. Closeness, like Degree, only considers direct connections and so is a local measure. Closeness can help find ‘broadcasters’, however in a highly connected network, many nodes often have a similar score and thus it has better use in finding influencers within a single cluster.

***Figure 4****: Graph network coloured by centrality value. Image sourced from: https://en.wikipedia.org/wiki/Centrality*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Centrality Measure** | Number of nodes | Fellowship  development | Assortative mixing | Shortest Path | Structure of  network | Position | Dissimilarity |
| Degree |  | ✖ | ✖ | ✖ | ✖ |  | ✖ |
| Betweeness |  | ✖ |  |  | ✖ |  | ✖ |
| Closeness |  | ✖ | ✖ |  | ✖ |  | ✖ |
| Eigenvector |  |  |  | ✖ |  |  |  |

Eigenvector centrality **(C)** calculates the local importance of a node within a specific group in a social network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Eigenvector centrality can be considered as a recursive version of degree centrality and as a result, this centrality measure takes the whole network topology into account.

***Table 3****: Characteristics of traditional centrality measures.* (Dhinesh Babu L.D)

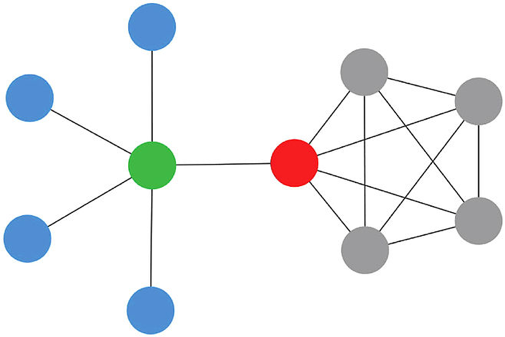
In all of these measures, the number of nodes a particular node is in contact with plays a crucial role in calculating its centrality value. In fact, a node’s centrality is directionally proportional to the number of connections the node has with other nodes. Another point to note is that the metric ‘position’ is not important here as there is no hierarchy in online social networks as any node will be able to communicate with another node which it is connected.

After reading numerous literature, it is certain that the centrality of a particular point cannot be assessed in isolation from the centrality of the other points to which it is connected. A point that is connected to more central points has its own centrality boosted, and this, in turn, boosts the centrality of the other points to which it is connected. There is therefore an inherent circularity involved in the calculation of centrality and the diversity of topological connections among the neighbours has a significant effect on a node’s influence. (Bonacich, 1972)

Assortive mixing is a property of bias in favour of connections between nodes with similar number of connections. It is known that social networks exhibit first-order assortive mixing; if two nodes are connected, they tend to have similar node degrees, suggesting that people tend to mix with those of comparable prominence. Second order assortive mixing is a feature which is almost unique to social networks. If two nodes are connected, we measure the degree correlation between their most prominent *neighbours*, rather than between the nodes themselves. This suggests that if two people interact in a social network, then the importance of the most prominent person each knows is very likely to be the same. Only Betweeness and Eigenvector centrality exhibits this property from the measures we have looked at. (Shi Zhou, 2017)

Fellowship development is defined in the original paper as “*the strong liking to grow association with other nodes in the network*”, however there is very little explanation about how this relates to the FBCS algorithm. Moreover, there is a contradiction in the original paper about whether Eigenvector centrality exhibits the characteristic of fellowship development. Due to lack of clarity in the original paper, it is necessary to study related literature in order to develop my own interpretation. Cuzzocrea et al hypothesise that “*stronger social identity* (i.e. higher levels of centrality) *leads to stronger desires to participate in the virtual community*” (Utpal M. Dholakiaa, 2004). This follows the nature of Eigenvector centrality where a connection to a high-scoring node contributes more to the score of the node in question than an equal connection to low-scoring nodes. As a result, there is a desire to develop links to nodes which are well connected, i.e. fellowship development.

Although not discussed in the original paper, I have included a final column labelled ‘dissimilarity’ as I believe it is an important feature of social networks and one which must be addressed in order to obtain better results in the centrality ranking of nodes in an OSN.

 In Figure 5, the green and red nodes are the most ‘dissimilar’ because they do not share neighbours between them. As a result, the green node contributes more to the centrality of the red one than the grey ones because the red node can access the blue ones only through the green node. Similarly, the grey nodes are redundant for the red one because it can access each grey node without any intermediary.

This measure permits us to quantify the topological contribution of each node to the centrality of a given node, having more weight for those nodes with greater dissimilarity since these allow the given node access to nodes that which themselves cannot access directly (Alvarez-Socorro, 2015). The only traditional measure which takes this into account is Eigenvector centrality which is calculated through solving the eigenvalue problem: such that where A is the adjacency matrix of the network and D is an arbitrary dissimilarity matrix.

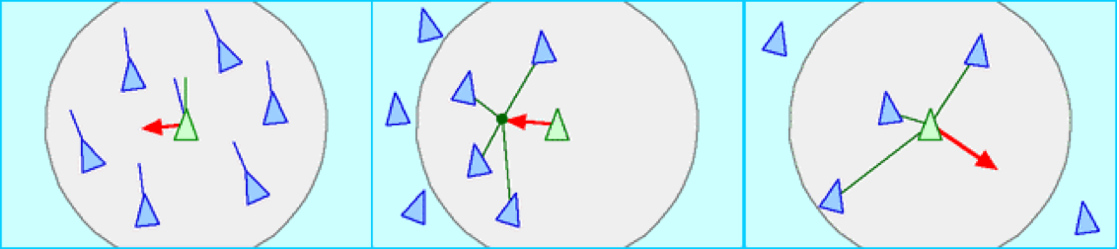
***Figure 5****: Dissimilarity based measures. Image sourced from: https://en.wikipedia.org/wiki/Centrality#Dissimilarity\_based\_centrality\_measures*

The new centrality measure proposed aims to exhibit assortive mixing, fellowship development as well as consider the number of connections a node has. As an extension, I aim to evaluate the effects that incorporating dissimilarity into the FBCS algorithm would have.

**F Boids**

The proposed algorithm which I am validating is inspired by the flocking behaviour of birds as they exhibit characteristics such as coordinated behaviour and their ability to change direction with the flock which, as we shall see, closely mimics the behaviour of humans in OSNs. Boids is an ‘artificial life program’ created by Craig Reynolds in 1989 which simulates this “evolutionary” or emergent behaviour; that is, the complexity of Boids arises from the interaction of individual agents following a set of local rules. Each boid (*simulated bird-like object)* behaves autonomously and gets local neighbourhood information and although the long-term behaviour of an entire flock is difficult to predict, its motion and arrangement is predictable and orderly over small periods of time. (Wong, 2008)

A flock is defined as “*a group of entities that demonstrate an aligned, non-colliding and aggregate motion*” (Reynolds, 1987). These three descriptors form the basis of the three rules, described below, that govern Boids algorithm which can be related specifically to OSNs and thus are used in developing the new centrality measure.

Firstly, birds in a flock change their direction and speed to avoid collisions with other birds in the flock. This rule is known as separation. Secondly, birds in a flock have the propensity to stick to the centre of the flock. This results in cohesion in the flock. Thirdly, flocking birds change their speed and direction relative to their neighbours to make the flock movement uniform. This velocity and direction matching results in alignment of the boids. These three properties make it possible for a flock to bind around a centre of mass and move in a smooth, uniform flow. One point to note is that all these rules are local; each boid is independent and interacts with others according to its local perception of the flock.

***Figure 6****: Illustrations representing Alignment, Cohesion and Separation properties (left to right). Image sourced from: Boids by Craig Reynolds, sourced from: https://www.red3d.com/cwr/boids/.*

OSNs can be compared to a flock of birds in many aspects. Users in an online social network stay away from people they disagree with which is analogous to the separation rule. Moreover, users in a social network tend to congregate with people they like and share interests with which is related to the second rule of cohesion. Finally, people tend to give credence to an idea which is shared by a majority of people which is comparable to the third rule of alignment. Once again, these three observations are local; each human in an OSN is independent and interacts with others according to its local perception of the network. Furthermore, like flocking, behaviour in an OSN is autonomous, dynamic and there is no one leader. The observations derived from the above rules are utilised to frame the new centrality measuring algorithm.

**G Flocking Based Centrality Measure**

Centrality in OSNs is unique as it is dependent on both the number of connections as well as the frequency of interaction between these nodes. This is clear in OSNs as two friends who interact often will spread an idea much quicker than two people who are connected but seldom interact. As information regarding number of interactions is often not given in a static network graph, it has to be calculated approximately. I will return to this point in the following section.

In terms of the architecture of the system, it is utilising a component-based architecture which decomposes the algorithmic design into three separate functional procedures where each procedure is based on one rule of Boids. As a result, this is a main program with subroutines where the components are the main program and the subroutines, the connectors are the procedure calls and the data is the values passed into and out of the subroutines. In terms of topology, the static organisation is hierarchical and the full structure mimics a directed graph. This has an advantage of modularity and the ability to manipulate or add further subroutines as long as interface semantics are unaffected.

**H Interpretation of Flocking Based Centrality algorithm**

Let us denote our social network as the graph where V is the set of nodes and E is the set of edges. A “connection” is an edge of E (i.e. a direct link between two nodes in V) and an “interaction” reflects a path joining two nodes which are not necessarily connected. A node is a ‘neighbour’ to node if there is a connection between the two nodes.

1. **Interaction procedure**

The average interaction value of each node is calculated on the basis of the number of direct connections a node has. Let us denote the neighbourhood of a node by ; that is, is the set of nodes adjacent to with size equal to the degree of . Initially we compute the sum of the degrees of the nodes in , denoted , which is recomputed as /|so that is now the average sum of the degrees of the nodes in Taking some neighbour of , we can consider the degree of to be a measure of the amount of ‘gossip’ that will come into from . Hence, the sum of the degrees of the neighbours of is a measure of the overall amount of ‘gossip’ that will come into from its neighbours. If we divide the sum of the degrees by the number of neighbours then the resulting value is a measure of the average amount of gossip that will come into from one of its neighbours. We subtract the number of neighbours from this value as the more neighbours of , the greater the reduction of . The lessening of corresponds to the loss of influence of a neighbour as follows. Suppose has 3 neighbours and has 100, where the values for and , and , say, are identical. The revised values, and , say, are such that is greater than . This reflects a mix of the average amount of gossip per neighbour and the ‘influence’ of this information; is only influenced by 3 neighbours whereas is influenced by 100; hence, on average, a neighbour of v has more influence on than a neighbour of does on . The value obtained is divided by a scaling factor so as as to adjust the weight one chooses to give to the Interaction component.

1. **Separation procedure**

The Separation value can be considered to be a function of both the degree of a node as well as the number of interactions a node possesses. Let for ≥ 0, called the ball of distance around , be the set of nodes at a distance of at most from ; that is, for which a shortest path to has length at most This distance can be considered the ‘range of interaction’ and reflects where the ‘disliked’ nodes lie and thus we can say that a given node ‘interacts’ with if it lies in . We consider the sum of the degrees of the nodes of divided by the size of to give the average weight of gossip resulting from the nodes that does not dislike. The higher this value, the less ‘separated’ is within the network. This value is divided by a scaling factor so as as to adjust the weight one chooses to give to the Separation component.

1. **Relative vector / Alignment procedure**

We assume that a node ‘directly follows’ all of the nodes that it ‘likes’; that is, we assume that there is an edge from to every node in We then follow the same process as in the Interaction procedure to calculate the Cohesion value. One important design factor is the degree of separation which is the number of connections through which any two unknown users can be made to meet. The original paper uses the value of 6 as a scaling factor following Grigyes Karinthy’s work on the ‘6 Degrees of Separation’, particularly relevant to OSNs. The obtained value is divided by a further scaling factor so as as to adjust the weight one chooses to give to the Alignment component.

**I Evaluation**

The overall purpose of my project is the validation of the proposed centrality measure and so a rigorous evaluation strategy is essential. Evaluation of any centrality measure is a difficult task as there is no real indication of whether or not it is a ‘good’ measure. My evaluation will therefore be split into four sections.

Firstly, in relation to requirement NFR4, I will evaluate the validity of different centrality measures by conducting an empirical study analysing the correlation of various network centralities using real-world data from published network data sets. The correlation between pairs of centrality measures can be computed with the help of a correlation coefficient such as Kendall’s tau. A high correlation of the centrality measure would imply that the centrality measures are superfluous. On the other hand, if the measures are not highly correlated, they suggest that each of them is a distinct measure with different results.

Secondly, due to the algorithmic nature of the project, there is a strong emphasis on non-functional requirements, most notably in terms of efficiently. As a result, in relation to requirement NFR3, I will analyse and compare the new centrality measure to traditional ones in terms of computational complexity. This is extremely important as social networks can be composed of up to millions on nodes and edges and so the algorithm needs to be able to efficiently analyse these networks. I will also seek to evaluate the success of NFR1 and NFR2, analysing how changing parameters in the algorithm or adding a further procedure to account for dissimilarity between network nodes affects obtained values and the efficiency of the algorithm.

Thirdly, in relation to NFR5, as the data-sets are known industry-standard networks, it is possible to utilise existing knowledge from associated published literature to review the role of each centrality measure in the determination of influential nodes and compare this to obtained results.

Finally, I will analyse my own work by comparing my obtained results using my implementation of the FBCS algorithm to the results obtained in the paper for each benchmark data-set.

# References

Alvarez-Socorro, H.-A. G.-D., 2015. Eigencentrality based on dissimilarity measures reveals central nodes in complex networks. *Universidad de Chile, Santiago, Chile..*

Anon., n.d.

Barabási, R. A. a. A.-L., 2002. Statistical mechanics of complex networks. *REVIEWS OF MODERN PHYSICS,* Volume 74.

Bonacich, P., 1972. Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology.*

Borgatti, S. P., 2005. Centrality and network flow. *Social Networks 27.*

D. Lusseau, K. S. O. J. B. P. H. E. S. a. S. M. D., 2003. The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology.*

Danah M. Boyd, N. B. E., 2007. *Social Network Sites: Definition, History, and Scholarship.* [Online]   
Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/full

Dhinesh Babu L.D, E. D. R., 2015. Flocking based evolutionary computation strategy for measuring centrality of online social networks. *ScienceDirect.*

Disney, A., 2014. *KeyLines FAQs: Social Network Analysis.* [Online]   
Available at: https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/

Girvan, M., 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences.*

Glance, L. A. A. a. N., 2005. The political blogosphere and the 2004 US Election. *WWW-2005 Workshop on the Weblogging Ecosystem .*

Hilgetag, M. K. a. C. C., 2006. Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS Comput Biol.*

Iyer, S., n.d. Limitations of graph centrality measures?. *Quora.*

Jalalzadeh, S., n.d. *Which open source software is best for network data analysis?.* [Online]   
Available at: https://www.researchgate.net/post/Which\_open\_source\_software\_is\_best\_for\_network\_data\_analysis

James F. Morris, J. W. O. a. R. F. D., n.d. A random graph generation algorithm for the analysis of social networks. *Journal of Defense Modeling and Simulation: Applications, Methodology, Technology.*

M. E. J. Newman, D. J. W. a. S. H. S., 2002. Random graph models of social networks. *Proceedings of the National Academy of Sciences.*

Marcon, A. M. a. M., n.d. *Measurement and Analysis of Online Social Networks.* [Online]   
Available at: https://people.mpi-sws.org/~mmarcon/SocialNetworks-IMC.pdf

Mariani, H. L. a. M. S., 2017. Ranking in evolving complex networks. *Arvix.org.*

N/A, n.d. *Social Network Analysis: Theory and Applications.* s.l.:s.n.

Reynolds, C. W., 1987. Flocks, Herds, and Schools: A Distributed Behavioral Model. *Computer Graphics,.*

Ron Hagan, Y. F. a. J. B., n.d. Centrality Metrics. *University of Tennessee Knoxville.*

Scott, J., 1991. *Social network analysis.* s.l.:Sage.

Shi Zhou, I. J. C. L. K. H., 2017. Second-Order Assortative Mixing in Social Networks. *Cornell University Library .*

Stephen P Borgatti, M. G. E. J. C. J., n.d. *Analyzing Social Networks.* s.l.:SAGE Publishing .

Steve Borgatti, L. K. T. V., n.d. *International Network for Social Network Analysis.* [Online]   
Available at: http://www.insna.org

Strogatz, D. J. W. &. S. H., 1998. Collective dynamics of ‘small-world’ networks. *Nature, International Journal of Science.*

Thomas W. Valente, K. C. C. L. a. E. C., n.d. How Correlated Are Network Centrality Measures?. *US National Library of Medicine.*

Travis Martin, X. Z. a. M. E. J. N., n.d. Localization and centrality in networks. *Arvix .*

Utpal M. Dholakiaa, R. P. B. L. K. P., 2004. A social influence model of consumer participation in network- and small-group-based virtual communities. *Internation Journal of Research in Marketing.*

Wong, T., 2008. *Boids.* [Online]   
Available at: https://cs.stanford.edu/people/eroberts/courses/soco/projects/2008-09/modeling-natural-systems/boids.html

Wong, T., n.d. *Flocks of birds can be modeled with boids.* [Online]   
Available at: https://cs.stanford.edu/people/eroberts/courses/soco/projects/2008-09/modeling-natural-systems/boids.html

Xiaohui Zhao, F. L. J. W. a. T. L., 2017. Evaluating Influential Nodes in Social Networks by Local Centrality with a Coefficient. *Geo-Information.*

Zachary, W. W., 1977. An Information Flow Model for Conflict and Fission in Small Groups. *The University of Chicago Press Journal.*

1. A scale-free network is a network whose degree (number of direct connections to a node) distribution follows a power law. [↑](#footnote-ref-1)
2. Definition: G(n, p) is a random graph with n vertices where each possible edge has probability p of existing. [↑](#footnote-ref-2)