A Flow Propagation Method For Detection of Local Community

Saad Q. Albawi (Lecturer) * Hadeel T.Ibrahim (Lecturer)*

Tareq Mohamed (Asst.Lecturer)**

Abstract

This paper is using an algorithm (Flow-Pro) for finding the node community in a complex network without need to know the information of the whole graph. In general, the researchers supposed their network based on undirected graph and the edge weight for each two connected, neighbour nodes are equal to 1, otherwise it will be 0. In the first step, the function implemented to give community, according to the stored flow. Synthetic data were used with 20,000 nodes. Also, 20 communities had been used. In this paper, edges weights N x N for network used, where N denotes the number of nodes. The total number of messages that produced from the flow algorithm for 1000 nodes was calculated (299392), where for 20000 nodes in our result was (45,582,924) messages.

Keywords: Flow Propagation, Local Community, Social Networks, Community Detection.

*Diyala University

**Kirkuk University

1. Introduction

Community is a group of individuals (nodes) in a shared social media that can interact with each other by their common ideas, interests, jobs, etc. [1]. Local communities, are densely-connected node sets that discovered and evaluated based only on local information [2]. Community detection is an important subject in social networks, but it contains many obstacles. There are a lot of community detection applications, for example, finding web communities is one of them. Other community detection applications are detecting the structure of social networks, analyzing a graph's structure to uncover Internet attacks and image segmentation. These applications are the most important application. the Flow-Pro algorithm is used in this paper to detect local communities and we applied this algorithm with our own code using Matlab 2014a as a platform.

2. Related works

Some previous works, detected communities by using a specific mining measure named Max-Min Modularity which mentioned the connected pairs and the defined criteria for each pair [3]. Different approaches are generally described as "community discovery" that was made to provide a formal definition of the concept. Another work based on Web Self-Organization even the nodes structure was decentralized and unorganized, in this paper they found communities using the connectivity information only [4]. Variations also appeared in the method used to identify the community. Some algorithms uncovered entire network, or each node and division in the community or merge them [3] respectively, communities, producing a hierarchical tree called nested communities. Many researchers are aiming to find entire hierarchical where others only want to define the optimal community section [5] or by uncovering the whole structure of the network [6].

3.Problem definition

Suppose there are nodes in a network and some nodes for the community. The initial node sϵV (G = (V, E)) is given. To find the community of s (C(s)), V $\supseteq$ C(s) $\supseteq$ {s} so there exists high number of edges between the nodes of C(s) comparing with the number of edges that connects the nodes of C(s) and the rest graph. Let p (x), x $\in$ V denotes the probability that node x belongs on C(s). Also, Let d (x), x $\in$ V denotes the shortest path distance between the nodes x and s. A simple estimation of p(x) can be given by Equations (1) and (2) .

$$p(x) = \rho^{-d(x)} \quad (1)$$

$$p(x) \leq \frac{\sum_{y \in n(x)} p(y)}{|n(x)|} \quad (2)$$

Where ρ be the average ratio of local links to node degree value and n(x) denote the set of neighbors of node x

4. Motivation

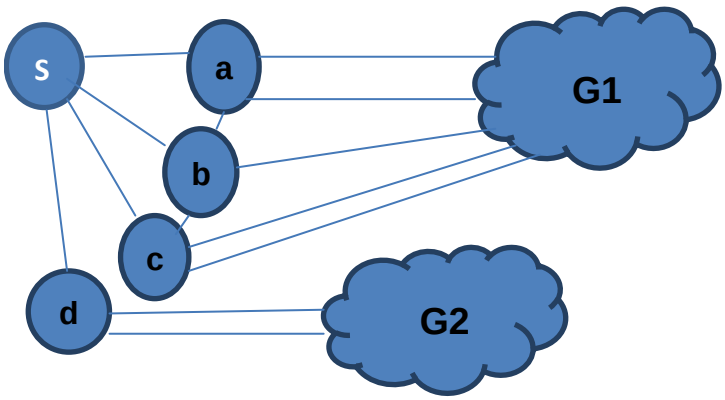
The motivation in our paper is appeared clearly by using the Flow-Pro algorithm to detect local communities without using the information for the entire graph. Algorithm is coded mostly in different way and discovering the local communities in social networks inefficient approach by calculating the stored and transmitted flow for each node may be belong to the specified community.

5. Data Collection

The current paper simulates the data randomly. It used rand function for creating the random data between 0 and 1. The two data sets resulted were saved in database files (data1000.mat for 1000 nodes and data20000.mat for 20000 nodes).

6.Methodology

Flow-Pro algorithm is used in each one of the main processes, emits a stream shared the first node neighbor [7]. Each node stores a flow to spread it to its neighbors and able to return a part of the flow to the first node. The p(x) is analogous on stored flow of node x. There are four phases of the algorithm. Figure (1) shows a sample for community network.



Figure(1) Community Network

First Phase: In each iteration of the main process, a node emits a stream shared by the adjacent edges of the first nodes weight. Each node stores a stream half area (S(x)) is the stored flow which is actually equal to the half flow and transmits the other half T(x) to the neighbor node). The stored flow, S(x) should be less than a threshold value in order to end the flow. This process is based on the equation (1) hereinafter; p nearby nodes will be stored in the high flow. In addition, the Flow-Pro algorithm, consider the importance of the node to belong to C(s), where in previous approaches, they based on the value of d (x) only, which means the shortest path between x and s nodes. The importance of the node in our approach calculated by the quantity of the stored flow S(x) for each node, the importance increased when the S(x) is increased also.

Second phase: Proposed method removes and adds nodes in the current phase by considering the stored flow for each node. The extension for lifting purpose to decrease shortest paths between the nodes that belong on the community of s and s in order to be able to increase their stored flow in the next iterations, to gradually keep the most of the flow to the nodes of the community by removing bridges and to keep the number of neighbors of s balanced.

Third phase:

$$p(x) \leq \frac{\sum_{y \in n(x)} p(y)}{|n(x)|}$$

Based on this equation, in the case that S(x) is greater than E(S(n(x))), the researchers set it to E(S(n(x))). This step will clearly decrease the S(x) from nodes that does not belong on the community.

Fourth Phase : There exists a bridge (edge: s ~ d in Figure (1)) and due to the third phase the reduction of S (d) will be high. Without this step, S(d) will be less but close to S(a), S(b) and S(c). The vector S is sorted in descending order and the differences are computedbetween adjacent elements of the sorted vector DS. Let K be the position of the global minimum of DS. The community of nodesis defined by the first K nodes with highest S (x). The community finding algorithm converges to a solution (e.g. the last 10 iterations that receives the same community). The quantity $\frac{T(s)}{\sum_{x \in V} S(x)}$ is less than the specified threshold which leads to that S(.) has been converged.

7.Results

In the first step, the functionimplements and produces communities according to the stored flow. Figures (2 and 3) show the proposed networks for communities with 20000 and 1000 nodesrespectively. The communities grouped in more clear way in figure (3) than figure (2) because the number of nodes in figure (3) is less than the number of nodes in figure (2).

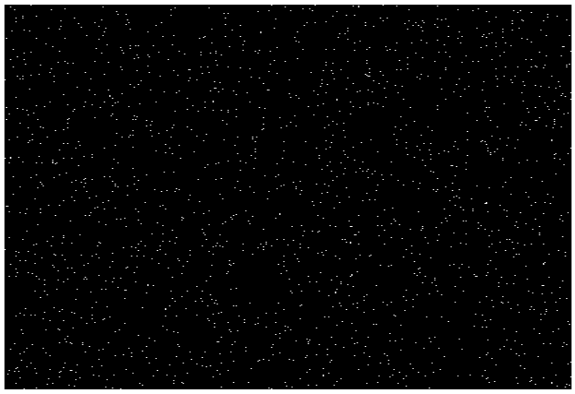


Figure (2). The proposed network for 20000 nodes, white grouped dots referred to communities.

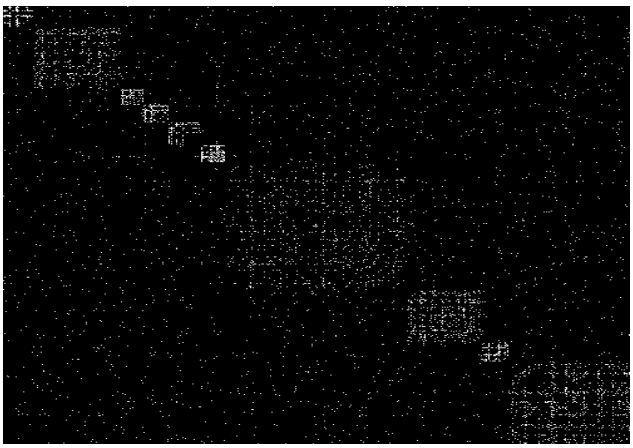
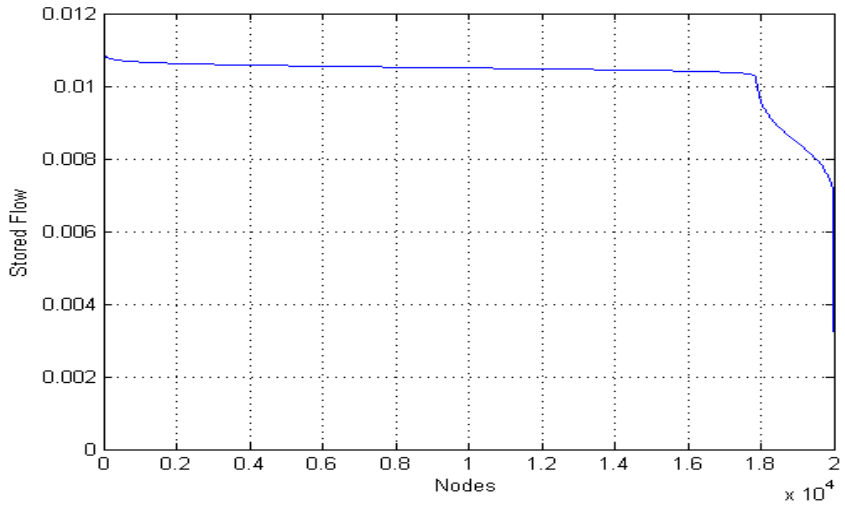


Figure (3). The proposed network for 1000 nodes, white grouped dots referred to communities.

In figure (4) the stored flow is relatively high in the first 18th thousands nodes and suddenly decreased because they are not neighbors to s node. In 1000 nodes network (Figure (5)), the stored flow reduced after the 50th for the same reason. Figures (6 and 7) shows the stored Cluster (for both 20000 and 1000 Nodes respectively). Figures (8 and 9) shows the curvesfor thearray of communities for (20000 and 1000 nodes respectively). Figures (10 and11) show the edges weights for (20000 and 1000 nodes respectively) where (1 for blue, 0 for white). While figures (12 and 13) explain theInitial weights in each iteration for (20000 and 1000 nodes respectively). Finally figures (14 and15) explain sending initial weights in each iteration for (20000 and 1000 nodes respectively).



Figure(4) Stored flow vs. Nodes for 20000 nodes.

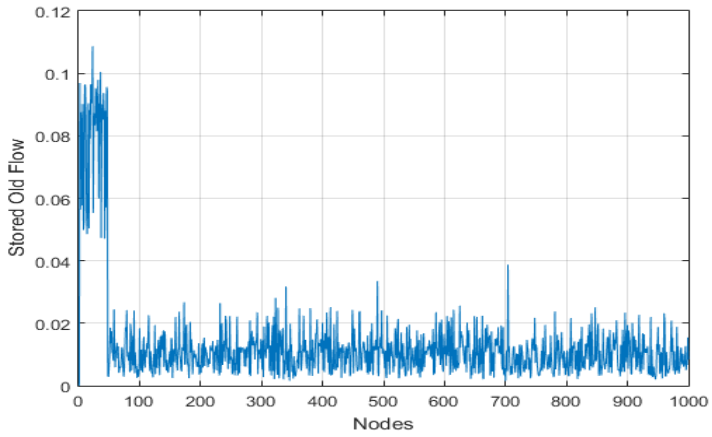
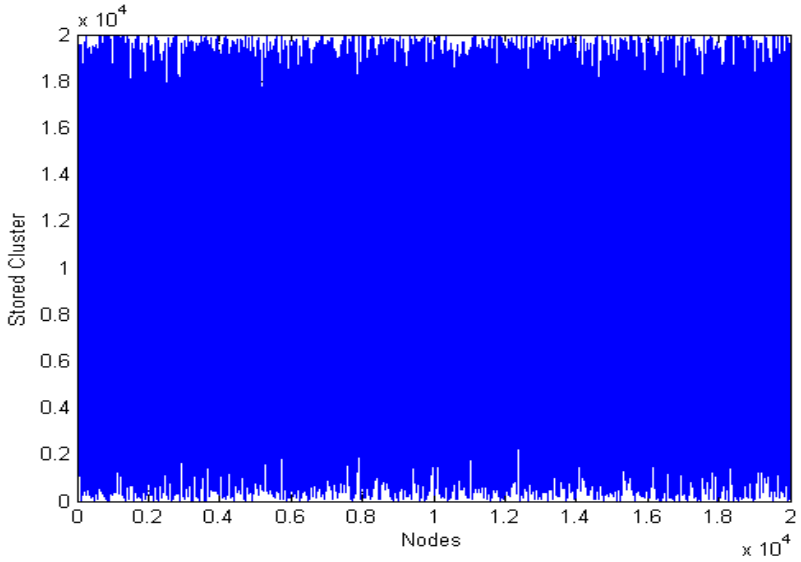
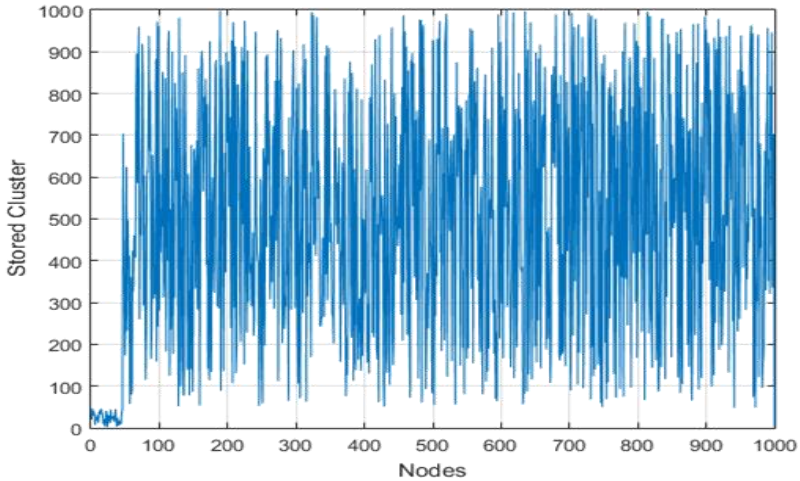


Figure (5) Stored flow vs. Nodes for 1000 nodes.



Figure(6) Stored Cluster vs. Nodes for 20000 nodes.



Figure(7) Stored Cluster vs. Nodes for 1000 nodes.

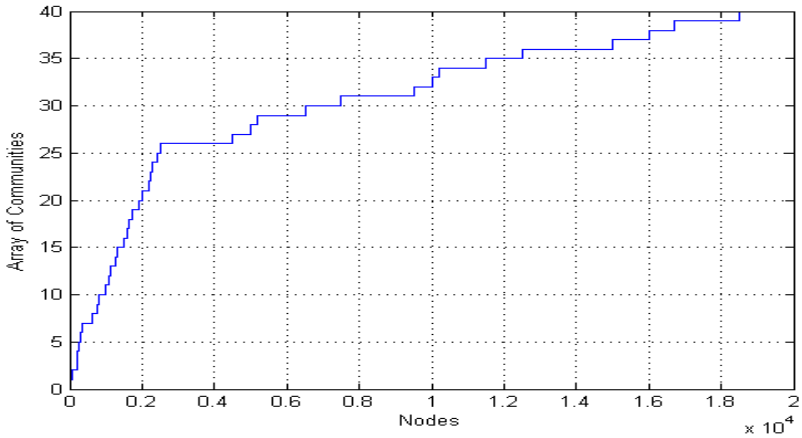


Figure (8). Array of communities vs. nodes for 20000 nodes.

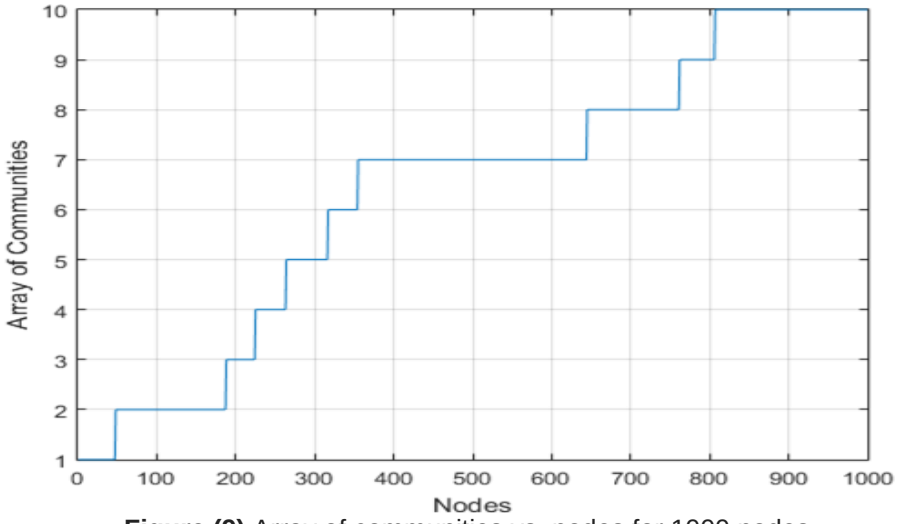


Figure (9) Array of communities vs. nodes for 1000 nodes.

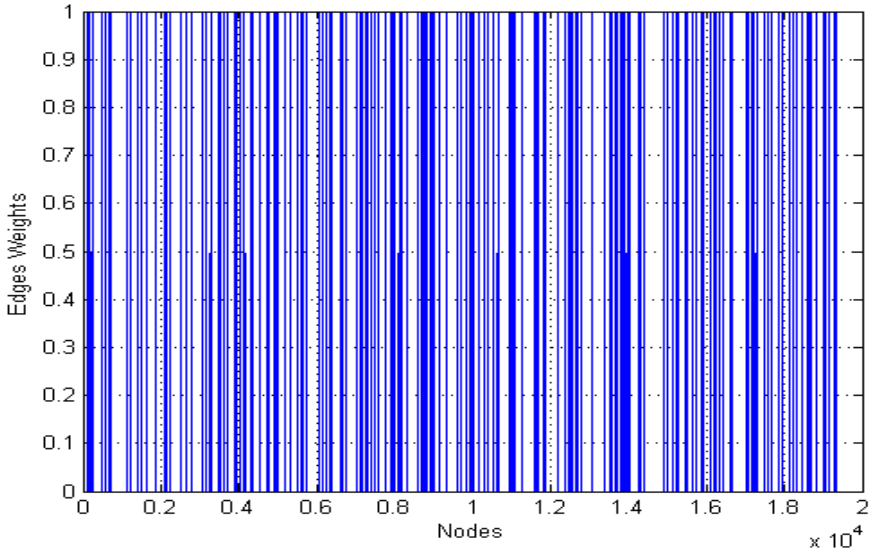


Figure (10) Edges weights vs. nodes for 20000 nodes (1 for blue, 0 for white).

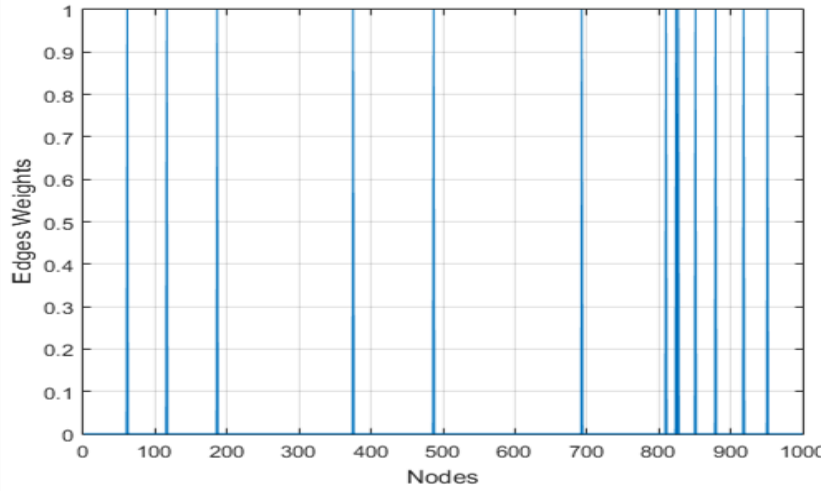


Figure (11) Edges weights vs. nodes for 1000 nodes (1 for blue, 0 for white).

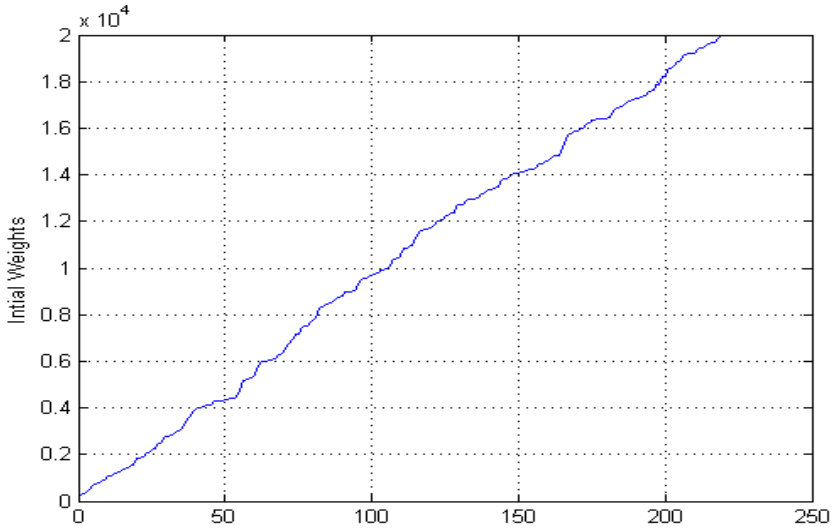


Figure (12) Initial weights in each iteration for 20000 nodes.

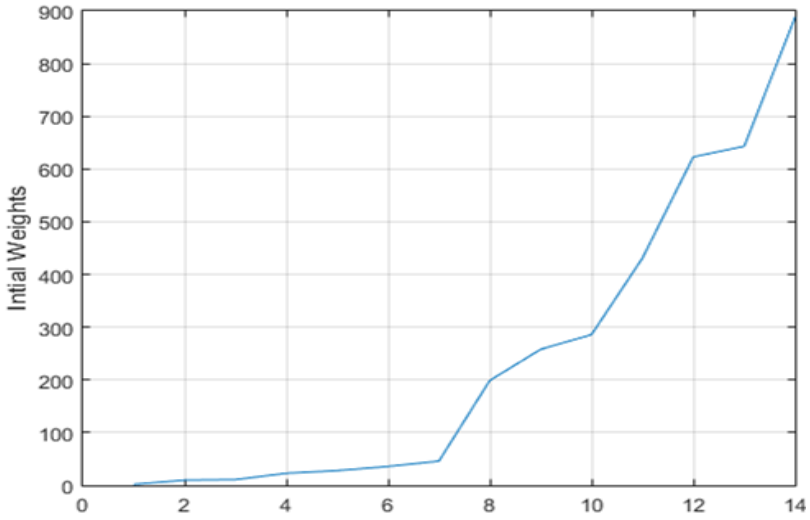


Figure (13) Initial weights in each iteration for 1000 nodes.

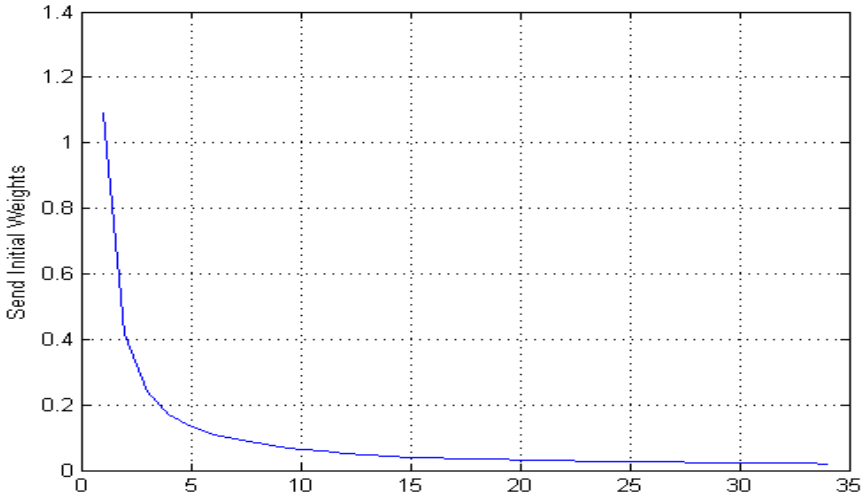


Figure (14) Sending initial weights in each iteration for 20000 nodes.

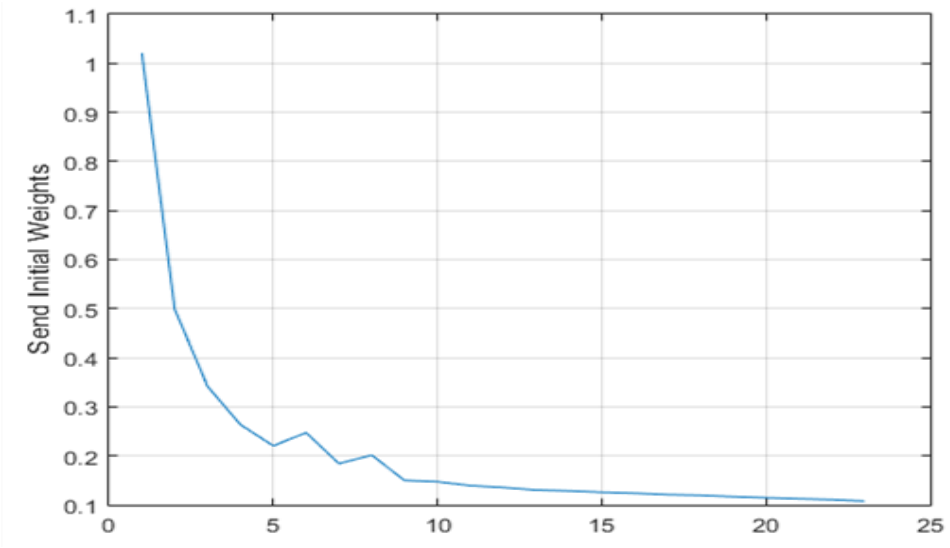


Figure (15) Send initial weights in each iteration for 1000 nodes.

8. Conclusions

In this paper the researchers worked on an algorithm for finding the node community in a complex network without the information of the whole graph that is why it will be local and differs from the existing methods from literature. In the first step, the function has implemented to identify the community according to the stored flow. Synthetic data used with 1000 nodes, also 20 communities were used. In this paper, the edges weights N x N are used for network, where N denotes the number of nodes. The Number of iterations were calculated which needed for flow algorithm. At the end, the number of iterations is calculated which is needed for flow algorithm. The total number of messages of flow algorithm for 1000 nodes was calculated (299392) but for 20000 in the result is (45,582,924).

9. Recommendations

In future, researchers can use another scenario for finding the optimum community structures. Also they can test this method on real community network. The plan is to expand the algorithm to ensure that the perception of overlapping and non-overlapping communities.

References

[1]http://en.wikipedia.org/wiki/Virtual_community, March, 2011.

[2] A. Clauset, " Finding Local Community Structure in Networks", Pysic.data-an, pp.1, Feb. 2, 2008.

[3] J. Chain, O. R., Randy G., "Detecting Communities in Social Networks using Max-Min Modularity", PP.1, May, 2009.

[4] D. Katsaros, G. Pallis, K. Stamos, A. Vakali, A. Sidiropoulos, and Y. Manolopoulos, "CDNs content outsourcing via generalized communities," IEEE Transactions on Knowledge and Data Engineering, vol. 21, pp. 137–151, 2009.

[5] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self organization and identification of web communities," IEEE Computer, vol. 35, pp. 66–71, March 2002.

[6] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," Physical Review E, vol. 69, no. 2, p. 026113, Feb 2004.

[7] Costas Panagiotakis, Harris Papadakis, and Paraskevi Fragopoulou, "FlowPro: A Flow Propagation Method for Single Community Detection", IEEE 11th Consumer Communications and Networking Conference, Jan., 2014.

طريقة تدفق الانتشار للمجتمع المحلي

م.سعد قاسم فليح* م.هديل طارق ابراهيم* م.م.طارق محمد**

المستخلص

في هذه المشروع استخدم خوارزمية (تدفق برو) للعثور على عقدة الأجتماع في شبكة معقدة من دون الحاجة إلى معرفة المعلومات من الرسم البياني كله. بشكل عام تم افتراض شبكة تعتمد على الرسم الغير مباشر ووزن الحافة لكل عقدتين متجاورة متصلة يساوي (1)، وإلا فإنه سيكون (صفر). في الخطوة الأولى تم تنفيذ الدالة التي تعطي الأجتماع بالاعتماد على مسار مخزون البيانات. استخدمت بيانات افتراضية مع (20000) عشرون الف عقدة وكذلك باستخدام (20) عشرون من الاتصالات. في هذا المشروع استخدم اوزان الحواف (س * س) للشبكة حيث ان (س) يمثل عدد العقد في الشبكة. العدد الكلي للرسائل التي ولدت من خوارزمية التدفق ل (1000) عقدة كانت (299392) رسالة. بينما عدد الرسائل ل (20000) عقدة كانت (45,582,924) رسالة.

* جامعةديالى
**جامعة كركوك