ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 133
2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE2016)

A Definition of Structure Hole and Algorithms

Tianji Zhao
Tinghua University, Beijing 10084, China

***Abstract*—In this paper we only use graph's structural information to generate other information in order to get the structural hole. Our work firstly delivers a formula well defines the structural hole, gives an iterative algorithm and a descending algorithm to find solution. The definition combines the advantages of constraint method and two neighbors' algorithm, and avoid their disadvantages. Roughly speaking, the definition that determines the tie's weight that includes global information and uses local parameters to avoid confusion between opinion leaders. For the reason that nodes' and edges' weight are loop definition, the fixed point of the loop is the optimal solution. We design two algorithms to find the solution, analysis their convergence, correctness. The chosen of the initial point in iteration algorithm cases its divergence. The random steepest algorithm has a good property that it converges at any case although with a slow convergence rate. The new point in this paper is that we deliver an new definition and the corresponding method**

Keywords-component; Structure hole; fixed point; interated algorithm and descending algorithm

I. Introduction

Social network analysis often focus on macro-level models such as degree distributions, diameter, clustering coefficient, communities, small world effect, preferential attachment, etc.

We discuss structural holes of social network in this literature. Online network enlarges our daily life and provides us various opportunity to make friends with unfamiliar people. Social network theories mostly focus on analyzing the relationship between entities, the influence occurred by the structure, the impact of the individual's role over the network.

Research on relationship could be classified into different categories by different benchmarks: strong tie versus weak tie, one-way relationship versus two-way relationship.

The one-way relationship is often a relationship as fans in Weibo. Two-way relationship in other hand is typically a friendship between two people such as friends in Facebook. "Mass Communication and Para-social Interaction"[4] is a inchoate work focusing on the difference of the one-way and two-way relationship using the television as example. "Social influence analysis in large scale networks"[7] illustrate the influence of topic between individuals on the social ties globally and quantifies the relationship. Research on strong, weak social ties concerning with social capital depends on the job searching problem. In "How college affect students"[6] social network could benefit student through information transformation. The later work "Social Capital and Finding a Job: Do Contacts Matter"[5] suggests that the social capital mostly make similar people to be friend.

The structure hole is a subpart of the social capital in some sense. In "Structure hole" [1], Burt gives an illustration about structure hole. It is a gap of information transformation between the different communities or closures. "Strength of Weak Ties"[3] focuses on the strength of weak ties in the structure hole, "A set of measures of centrality based on betweenness "give a concrete definition of structure hole by betweenness centrality. "Structural Holes: The Social Structure of Competition" [2] considers structure hole in the network with redundancy as the cost of information and energy transformation. "Strategic leadership network brokerage" measures the structure hole using constraint model, which weighing the resource concentration in directed and undirected case.

II. Model

A. Our Model and Following Properties

Because

$$(\sum_{k\in(N(i)\bigcap N(j)),j} p_{ik}p_{kj})^2 \leqslant (\sum_{k\in(N(i)\bigcap N(j)),j} p_{ik}^2)(\sum_{k\in(N(i)\bigcap N(j)),j} p_{kj}^2) \leqslant (\sum p_{ij}^2)^2$$

The last equation hold only when p's value distributed equally. The disadvantage of the above model is that we could not distinguish the isolated node with the structural hole.

Our model is defined as follows:

- G is the graph of network with the node set N and edge set E
- C is a vector of node's weight
- P is a matrix of edge's weight

The node and edge's weight are defined as:

$$(\sum_{k\in(N(i)\bigcap N(j)),j} p_{ik}p_{kj})^2 \leqslant (\sum_{k\in(N(i)\bigcap N(j)),j} p_{ik}^2)(\sum_{k\in(N(i)\bigcap N(j)),j} p_{kj}^2) \leqslant (\sum p_{ij}^2)^2$$

The last part of the formula is $\frac{1+\#\ of(m,n),m,n\in N(i)}{\# of p_{mn}}$ which will increase if the node is the structural hole of the network and independent of j. This part has perfect properties that

1. all the nodes connected to the balanced cluster will have larger weight
2. the node connected to only one cluster has low value.
3. the result of the first round is that c_i with high constraint is larger than the normal one.
4. the disadvantage of the algorithm is that if two cluster coincide with each other on several certain node the c value will also be large.
5. the value of the edge will increase if and only if it connects to a high value node.
6. the iteration will maintain the structure of the graph that p_{ij} will be 0 if and only if it is 0 initially.

Proof:

1. The node connected to balanced cluster will have larger weight. The last part of the equation is only determined by the structural of the graph. Therefore by the equation (4) (5),

The node which connected to the cluster that has equally edge-values is larger than the node connected to unbalanced cluster.

2. The node connected to single cluster is lower than multiple clusters. The last part of the equation (7) is proportional to the k and the front is proportional to $\left(\frac{1}{k}\right)^2$. Generally speaking the coefficient is proportional to the $1/k$ where k represent the number of clusters the node belongs to.

3. If two cluster collapse with each other on several certain nodes the c, value will also be large. This property is similar to property (2) but the increase of the coinciding node number will lead to the decrease of c value. It is intuitive that c value will decrease because it becomes to be a new cluster if the coinciding number is large enough.

4. Value of the edge will increase if and only if it connects to a high value node. The increment of c in each iteration is determined by $\sum_{k\in N(i)\cap N(j)} P_{ik}P_{kj}$ The node connected to the node that has high value and various common neighbors.

5. The iteration will maintain the structure of the graph that p_{ij} will be 0 if and only if it is 0 initially

6. The identity part of the node $I(p_{ij} \neq 0)$ convince that the iteration will not change the structure of the graph. Maintaining the Integrity of the Specifications

III. Iterated Algorithm

A. *Existence of Soultion*

The best constraint value of the social network using this definition will be the fixed point of the algorithm, because the node has the properties listed in the above section. So what to do next is :1) simplify the question 2) discuss whether there exist a fixed point 3) make the iteration algorithm converge to the result.

$$c_i^{'} = \sum_j (c_i/\beta_j + \sum_k \frac{c_i c_k}{\beta_k \beta j})^2 \#(i) \qquad \beta_i = \sum_{k\in N(i)} c_k$$
$$= \sum_j [c_i/\beta_j (1 + \sum_{k\in N(i)\cap N(j)} \frac{c_k}{\beta_k})]^2 \#(i)$$

The main idea of our proof is that 1) each iteration of the algorithm is a function from a solution space into itself. 2) This function has a fixed point on its solution space. 3) In a finite iteration the result will converge to the fixed point.

We denote the solution space E as a set o a set of $n \times n$ matrix M which's each element $M_{\{ij\}} = p_{\{ij\}}, i \neq j$ with $M_{\{ii\}} = 0$. By the property that $p_{ij} = c_i/(\sum_{k\in N(j)} c_k$, $\sum_{i\in N(j)} M_{ij} = 1$ is trivial. Then we can use a function f to represent each interaction of P_{ij} into itself, specifically we have

$$f: f(M_0) = M_1$$

where $M_0(ij) = p_{ij} \backslash M_1(ij) = \frac{c_i}{\sum_{k\in N(j)} c_k}$

The solution space E is a subspace of the matrix space S with the Forensics norm $| A| = \left(\sum_{i=0}^{m} \sum_{j=0}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}$

With the above assumption, we analysis the convergence of the algorithm.

Theorem[Schauder]

Let K be a no-empty closed convex subset of a norm space. Let T be a continuous mapping of K into a compact subset of K. Then T has fixed point in K

The solution space E is a convex space, because the linear combination of the matrix in space E are still in E. For every node in space E, we have:

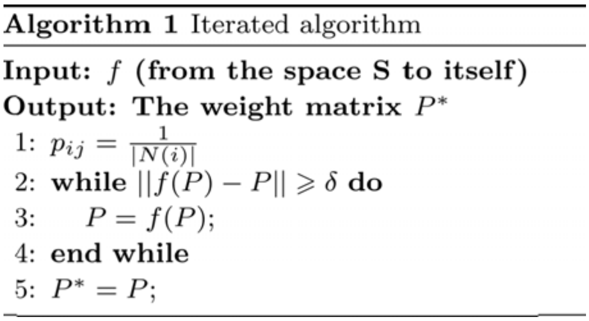
$$||M_t|| = ((\sum_{i=0}^{m}\sum_{j=0}^{n} |M_t(ij)|^2)^{1/2})$$
$$\leqslant (\sum_{i=0}^{m}(\sum_j M_t(ij))^2)^{1/2} = \sqrt{m}$$

Therefore the space E is bounded and the function is continuous. Thus, the function has a fixed point on the solution space. The fixed point in this solution space is a weight distribution of nodes and edges in the space with the perfect properties above. Hence the fixed point of the projection is the value of structural hole over the whole network.

B. *An Interated Algorithm*

How to find the fixed point of the projection in an executable algorithm comes to be important. In this part, we focus on the correctness, convergence and application condition of the algorithm.

We use the iterated way to calculate the fixed point:



Algorithm 1 Iterated algorithm

```
Input: $f$ (from the space S to itself)
Output: The weight matrix $P^*$
$p_{ij} = \frac{1}{|N(i)|}$
while $||f(P) - P|| \geqslant \delta$ do
    $P = f(P)$;
end while
$P^* = P$;
```

In the following paragraphs, we consider the influence of the algorithm's initialization condition and graph structure on its convergence.

Firstly the distance of the $d(f(M_A), f(M_A))$ is

$$d(f(M_A), f(M_B)) = (|\sum_i \sum_j (f(M_A) - f(M_B))^2|)^{1/2}$$

The convergence of the algorithm is equivalent to the convergence of the function (mapping) f. Therefore, we focus on the properties of the algorithm. We use the Edelstein theorem as follows:

Let T be a mapping of a complete chainable metric space(E, d)into itself, and suppose that there is a real number K with $0 \leq K < 1$ such that $d(x, y) \leq \epsilon \Rightarrow d(Tx, Ty) \leq Kd(x, y)$

Then T has a unique fixed point u in E, and $u = \lim_{n \to \infty} T^n x_0$ where $x_0$is an arbitrary element of E.

To analysis whether the function f is mapping function. We firstly give out the maximum increment of each element in the matrix under the case the total increment is less than

$$u = \lim_{n \to \infty} T^n x_0$$

To analysis whether the function f is mapping function. We firstly give out the maximum increment of each element in the matrix under the case the total increment is less than

With the assumption $d(M_A, M_B) \leq \varepsilon$, we indicate that

$$(\sum_{i=1}^{m} \sum_{j=1}^{n} |M_A(ij) - M_B(ij)|^2)^{1/2} = \epsilon$$

$$\Rightarrow \max\{|M_A(ij) - M_B(ij)|\} \leqslant \frac{\epsilon}{\sqrt{2}}$$

Proof: By the property of the matrix, we have $\sum_{i,j} M_A(ij) \sum_{i,j} M_B$. We denote $|M_A(ij) - M_B(ij)|$ as a_i (if$M_A(ij) \geq M_B(ij)$) and b_i (otherwise).We assume that the maximum point of $\max\{|M_A(ij) - M_B(ij)|\}$ is a series that $\sum_m a_m = \sum_n b_n = c$.For the reason a and b are independent, the maximum point must be the maximum of both a and b.

For a and b, if their summarization is a fixed number, their quadratic summarization reaches maximum when a and b are single number.

Therefore $\max\{|M_A(ij) - M_B(ij)|\} \leq \frac{\varepsilon}{\sqrt{2}}$

We give an assumption that the graph has k cliques. In order to find the upper bound and lower bound of the total incrimination, we consider $\beta = \sum c_i$ as.

$$\Delta\beta_i = \sum_{j \in N(i), \Delta p_{ij=0}} \{\frac{1 + \# \, of(m,n), m, n \in N(i)}{1 + \# of p_{mn}} [\Delta p_{ia} p_{aj}]\} \times$$
$$2(p_{ij} + \sum_{k \in N(i) \cap N(j)} p_{ik} p_{kj}) \;\; (p_{ia} \textit{ is the incremental edge})$$
$$+ \sum_{j \in N(i), \Delta p_{ij} \neq 0} \{\frac{1 + \# \, of(m,n), m, n \in N(i)}{1 + \# of p_{mn}} \Delta p_{ij} \times 2(p_{ij} + \sum_{k \in N(i) \cap N(j)} p_{ik} p_{kj})\}$$

We give an assumption that the graph has k cliques. In order to find the upper bound and lower bound of the total incrimination.

Lemma3.The increment of each element in matrix P is $-\frac{\Delta\beta_i}{\beta_i}$

Proof:

From $\sum_i a_i b_i \leq \sum_i b_i (\forall \; i, a_i \geq \; 0, b_i \geq \; 0)$, we know that $(\Delta \, p_{ia}) p_{\,aj} \leq \sum(\Delta \, p_{ia}) \sum p_{aj}$ **In the above proof, we use the conclusion $\Delta \, C_i/\beta = 0$ in inferring** $2\sqrt{2} \; \Delta\epsilon(1 + \#of(m, n), m, n \in N(i))/ \, (1 + \#ofP_{mn}) \,)$.

Lemma4: In balance network for each j, $\forall$ i, p_{ij} is equal, $\Delta \, C_i/\beta = 0$

proof:

$$\Delta c_i = \Delta \sum_{j \in N(i)} (p_{ij} + \sum_{k \in N(i) \cap N(j)} p_{ik} p_{kj})^2 \#(i)$$
$$= 2[\Delta \sum_{j \in N(i)} (p_{ij} + \sum_{k \in N(i) \cap N(j)} p_{ik} p_{kj})](p_{ij} + \sum_{k \in N(i) \cap N(j)} p_{ik} p_{kj}) \#(i)$$

$[\Delta \sum_{j \in N(i)} (p_{ij} + \sum_{k \in N(i) \cap N(j)} p_{ik} p_{kj})]$ is 0 when it is a balance clique(for each j,$\forall i, p_{ij}$ is equal).

Proposition:

The increment after mapping is big when the initial increment distribute concentrates to seldom, and small when it distributes averagely.

If the increment is on two edge, we denote the number of node as N and the number of clique as K

$$\Rightarrow \Delta\beta_i \leqslant 2\sqrt{2} \frac{1 + \# \, of(m,n), m, n \in N(i)}{1 + \# of p_{mn}} \Delta\epsilon / \sqrt{2} \times \sum_{j \in N(i) \cap N(j)} (p_{ij} + \sum_{k \in N(i)} p_{ik} p_{kj})$$
$$= 2\sqrt{2} \frac{1 + \# \, of(m,n), m, n \in N(i)}{1 + \# of p_{mn}} \Delta\epsilon$$
$$= 2\sqrt{2} \Delta\epsilon$$

If the increment is distributed averagely, we have

$$\Rightarrow \Delta\beta_i = \frac{1 + \# \, of(m,n), m, n \in N(i)}{1 + \# of p_{mn}} * \frac{\Delta\epsilon\sqrt{k}}{nN} \times 2 \sum_j (p_{ij} + \sum_{k \in N(i) \cap N(j)} p_{ik} p_{kj})$$
$$\leqslant 4 \frac{1 + \# \, of(m,n), m, n \in N(i)}{1 + \# of p_{mn}} \frac{\Delta\epsilon\sqrt{k}}{N}$$
$$= 4 \frac{\Delta\epsilon\sqrt{k}}{N}$$

Therefore the total increment is

$$\Delta = (\sum_i (\frac{\Delta\beta_i}{\beta_i})^2)^{1/2}$$
$$\leqslant \frac{4\sqrt{k}}{n} \Delta\epsilon$$

It is less than $\Delta \, \epsilon$ when k is large enough. Therefore, in this case, the node is converge to a fixed point.

Generally speaking, we can conclude that the mapping converge to a fixed point at the area where the increment is distributed averagely, the edge weight is distributed averagely and the local structure of each node is similar in each clique. An equivalent simplified condition is :(1)the local structural of each nodes is similar in each clique (the income edge and outcome edge's weight diverse little)(2) the initial edge weight of each node is 1 divide its degree. The reason is that if all the nodes are symmetric in each cliques, the increment incurred by the linking of two cliques will be transmitted from the connected node to the whole graph quickly and tiny increment will produce among the clique.

IV. Random Steepest Method

A. An Equivalent Represeting

We denote the iteration as a function from matrix P to P'. Generally speaking P'=f(P) has a fixed point when P=f(P) Therefore, we can represent the originally problem as

Finding P
P=f(P)

$$\text{s.t}\forall\ \text{i}, \textstyle\sum_{\text{j}} \text{P\_(ij)} = 1$$

From the above formula, we have several information :(1) The solution space is an convex space. (2) There exist at least one solution over the space.

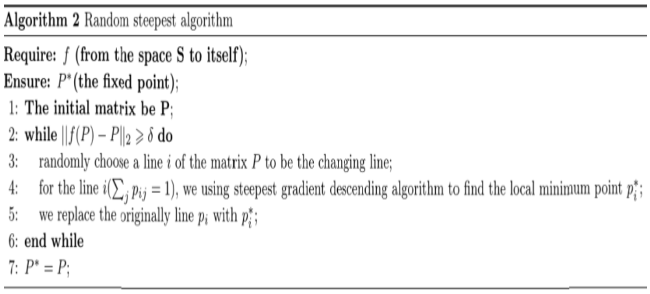
So we can change the above representation into an equivalent problem

$$\text{Finding P}$$
$$\min \left|\left|\text{f(P)} - \text{P}\right|\right|_2$$
$$\text{s.t.}\ \ \forall\ \text{i}, \textstyle\sum_{\text{j}} \text{P(ij)} = 1$$

An intuition is that the fixed point node is obviously the minimum point of the second problem. The solution of the second problem exists according to the convex theory. And it is equivalent to the first problem..

B. Typicle Method

We choose a line from the matrix in each round and fuse an advanced steepest algorithm.



Algorithm 2 Random steepest algorithm

```
Require: $f$ (from the space S to itself);
Ensure: $P^*$(the fixed point);
The initial matrix be P;
while $||f(P) - P||_2 \geqslant \delta$ do
    randomly choose a line $i$ of the matrix $P$ to be the changing line;
    for the line $i(\sum_j p_{ij} = 1)$, we using steepest gradient descending algorithm to find the local minimum point $p_i^*$;
    we replace the originally line $p_i$ with $p_i^*$;
end while
$P^* = P$;
```

The steepest algorithm we mention in the above algorithm is:

We denote the steepest descending direction as $(\beta_1, \cdots, \beta_n)$. At first, the concrete way is to find each $\alpha_{\text{j}} = (\Delta\ \text{g(P)})(\Delta\ \text{p}_{\text{ij}})$ /($\text{g(p)} = \left|\left|\text{f(P)} - \text{P}\right|\right|_2$). By the condition that the summarization of each line is 1. So finding the gradient direction is to solving the problem:

$$p_{ij}^{'} = I(p_{ij} \neq 0)c_j/(\sum_{k\in N(i)} c_k)$$

$$c_i = \sum_{j\in N(i)} (p_{ij} + \sum p_{ik}p_{kj})^2 \frac{1 + \#\ of(m,n), m,n \in N(i)}{1 + \#of p_{mn}}, \sum_{j\in N(i)} p_{ij}(0)$$

The reason why we use $\max\beta_0 \leq 1$ as an utilization condition instead of (the standard utilization condition for the derivation), is to simply the originals to a solvable problem. The gradient direction is $\text{d} = (\beta_0, \cdots, \beta)$.The above problem could use traditional convex optimization method to solve. The step length in this part is not strictly constrained. The main idea is to choose a length that decrease the function value as much as possible. Therefore $\lambda \geq 0$ is the smallest value with $\text{f}(\text{P} + \lambda\ \text{d}) \leq\ \text{f}(\text{P} + (\lambda + \Delta\lambda)\text{d})$ an d $\text{f}(\text{P} + \lambda\ \text{d}) \leq\ \text{f}(\text{P} + (\lambda - \Delta\lambda)\text{d})$

$$\lambda =\ \text{min\_}(\lambda \geq\ 0)\ \lambda\text{d}$$
$$\text{s.t}\ \text{f}(\text{P} + \lambda\ \text{d}) \leq\ \text{f}(\text{P} + (\lambda + \Delta\lambda)\text{d})\text{f}(\text{P} + \lambda\text{d}) \leq\ \text{f}(\text{P} + (\lambda - \Delta\lambda\ )\text{d})$$

In this part, we choose the minimum λ to guarantee that the function value decrease exactly which is a necessary condition in the following proof.

In the above algorithm, we give a descending algorithm to find the fixed point. In each iteration, we use bench gradient descending algorithm which will decrease the value of $\left|\left|\text{f(x)} - \text{x}\right|\right|_2$exactly. Therefore the correctness of the algorithm is equivalent to the convergence of the algorithm. Therefore we concern to the convergence again.

We firstly overview the proof of steepest descent method:

Lemma

If f(x) is continuous and differentiable, the solution set is $\Omega = \{\bar{\text{x}}|\nabla\ \text{f}(\bar{\text{x}}) = 0\ \}$, the sequence generated by the algorithm $\{\text{x}^{\text{k}}\}$ is contained by some tight set, convergence point of the sequence $\{\text{x}^{\text{k}}\}\ \hat{\text{x}} \in \Omega$

Proof:

we denote the algorithm as $\text{A} = \text{MD}$, which $\text{D(x)} = (\text{x}, -\nabla\ \text{f(x)})$. When $\text{d} = -\nabla\ \text{f(x)} \neq\ 0$, for M is close projection and f(x) is continuous and differentiable, D is continuous A is close at $\text{x}(\nabla \neq\ 0)$. For $\text{d} = -\nabla\ \text{f(x)} \neq\ 0$, f(x) is a decreasing function on Ω and A. For $\{\text{x}^{\text{k}}\}$ is in tight set, it is convergent. The proof of our algorithm is similar.

Theorem 2 The random steepest algorithm converge

We similarly consider the node $(\text{P}, \nabla\ \text{f(P)})$as a convert point.

We denote the algorithm as A=MD, which $\text{D(P)} = (\text{P}, -\nabla\text{f(P)})$. When $\text{d} = -\nabla\ \text{f(P)} \neq\ 0$, for M is close projection and f(x) is continuous and differentiable, D is continuous $\Longrightarrow$A is close at$\text{P}(\nabla \neq\ 0)$. The projection of our algorithm is also close for the reason that the differentiation of each line $\nabla\ \text{f}_{\text{i}}(\text{P})$is also bounded (every element is in [0,1]). For the reason that f(P) is decreasing on both two projection and $\{\text{x}^{\text{k}}\}$ is in tight set. So the algorithm converge.

Corollary 1The converge point is the locally minimum

The property of each step in random steepest algorithm converge algorithm is to find a new point with lower value. If the algorithm converge to a point, it means that it is the locally minimum point.

References

[1] Pierre Bourdieu and Loic Wacquant. An invitation to reexive sociology. The American Historical Review, 99(5):1644,

[2] Nald S Burt. Structural holes: The Social structure of competition. Journal of marketing, 58(1):152,1992

[3] Mark Granovetter. The strength of weak ties. American Journal of Sociology, 78(6):1360{1380, 1973

[4] Horton, Donald, Wohl, and R Richard. Mass communication and para-social interaction. Psychiatry MMC, 2015.

[5] Pascarella and Ernest T. How college a_ects students: Ten directions for future research. Journal of College Student Development, 47(5):508{520, 2006.

[6] Pascarellla and Ernest T. How college affects students: Ten directions for futhure research. Jouranl of College Student Development, 47(5):508-520,2006.

[7] Tang, Jie, Sun, Jimeng, Wang, Chi, and Zi Yang. Social inuence analysis in large-scale network