.....

*Keywords:* Local network, heterogenous network, node similarity

## 1. Local Network

A local network is a subset of network whose vertices are adjacent. The similarity of the nodes in Heterogenous networks is a new research direction. It is interested that how similar are these different vertices? or which node is most similar to others nodes?. In social network, an important people are connected with important friends. In other words, a nodes influence is not only decided by his contacts number but also decided by the influence of his local network on the whole networks. The influence is observed local networks around this node [**?** ]. In this research, a method which is based on the nodes local networks is proposed to quantify the nodes similarity in heterogenous networks. Each nodes local network is constructed by the directly connected nodes in the network. The details shown as follows:
The objective is to maximize the probability of neighborhood of a node $v$, denoted as $N(v)$, conditionally represented $f(v)$ where the probability of each node in the local network will be defined by probability of any node $c$ under $f(v)$ as

$$P_r(c)|_{f(v)} = \frac{\text{Node of interaction in LN}}{\text{Total degree of LN}}$$

Neighborhood of a node $v$ denied as

$$N(v) = N_1(v), N_2(v), ..., N_T(v)$$

where $N_t(v)$ denotes type $t$ neighbors of $v$ and $T$ is the number of node types. Here we call number of neighbors is the degree of node.
In this paper, a new methods which is based on Hellinger distance, a distance measure for probability distributions is proposed to describe the similarity of

those nodes in the heterogenous networks. The definition of the probabilities of each node is based on the degree distribution. Good news is that, measurement of similarity creates no problems; there are lot of similarity coefficients directed in the literature. Two completely similar objects give the maximum similarity usually $s_{i,i} = 1$ , while the least similar pairs reach the minimum value $s_{i,j} = 0$. That is, similarity is the complement of the dissimilarity measured in the range of $[0, 1]$, so one can be easily derived from the other:

$$S_{jk} = 1 - d_{jk} \tag{1}$$

or

$$S_{jk} = \sqrt{1 - d_{jk}} \tag{2}$$

### *1.1.* Heterogenous Network

Given a graph $S = (A, R)$ with set of entities types $A = \Lambda$ and a set of relations $R = \Re$, Heterogenous network is a directed or undirected graph $G = (V, E)$ with an object type mapping $\phi : V \rightarrow A$ and a link type mapping $\psi : E \rightarrow R$. Each object $v \in V$ relates to one particular object type $f(v) \in A$, and each link $e \in E$ relates to a particular relation $e \in R$. When the types of objects $|A| > 2$ or the types of relations $|R| > 2$, the network is called heterogeneous network; otherwise, it is a homogeneous network or bipartite network under the condition $|A| = 2$ and $|R| = 2$.

#### *1.1.1.* Theoretic definition of closed similarity

The previous similarity measures are tied to a particular application. Our goal is to provide a formal definition of similarity, we first clarify our intuitions about similarity.

**Intuition 1**: The similarity between $x$ and $y$ is related to their closeness.

**Intuition 2**: The more differences between $x$ and $y$, the less similar they are.

**Intuition 3**: When $x$ and $y$ are identical their similarity is maximum, no matter how much closeness they share.

We try to arrive at a definition of similarity that captures the above intuitions.

Since there are many alternative ways to define similarity. In this section, we first make a set of few assumptions about similarity that we believe to be reasonable. A similarity measure can then be derived from those assumptions. In order to capture the first intuition, we need a measure of closeness. Our first assumption is:

**Assumption 1**: The closeness between $x$ and $y$ is measured by

$$I(\text{closeness}(x, y))$$

where closeness between $x$ and $y$ is a proposition that states the closeness between $x$ and $y$; $I(s)$ is amount of information about $x$ and $y$. For example if $x$ is red and $y$ is orange then the closeness between $x$ and $y$ is color(x) and color(y). In information theory [**?** ], the information contained in a statement is measured by the negative logarithm of the probability of the statement. Therefore,

$$I(closeness(x, y) = -logP(color(x)\ and\ color(y))$$

Since knowing both the closeness and the differences between $x$ and $y$ means knowing what $x$ and $y$ are, we assume.

**Assumption 2**: The differences between $x$ and $y$ is measured by

$$I(descriptions(x, y)) - I(closeness(x)\ and\ (y))$$

where description $(x, y)$ is a preposition that describes what $x$ and $y$ are. From Intuition 1 and 2 the similarity between two objects is related to their closeness and differences. We assume that closeness and differences are the only factors.

**Assumption 3**: The similarity between $x$ and $y$, is a function of their closeness and differences i.e.

$$Sim(x, y) = f(I(closeness(x, y), I(description(x, y)))$$

the domain of $f$ is $(s, t), s \geq 0, y \geq 0, t \geq s$. From Intuition 3 when the two objects are identical the similarity measure reaches a constant maximum. We assume the constant is 1.

**Assumption 4**: The similarity between a pair of identical objects is 1. When

$x$ and $y$ are identical means knowing what they are,i.e., $I(closeness(x,y)) = I(descriptionof(x,y))$. Therefore the function $f$ must have the property $\vee s \geq 0$, $f(s,s) = 1$. When there is no closeness between $x$ and $y$ , we assume similarity is 0.

**Assumption 5**: $\vee\ t > 0, f(0,t) = 0$

Suppose it is possible to view two objects $x$ and $y$ from two separate perspectives. From each perspective, their similarity can be calculated separately. For example, by comparing the word sets in the documents or comparing their stylistic parameter values, the similarity between two documents can be calculated. We assume that the overall similarity between the two documents is a weighted average calculated from different perspectives of their similarities. The weights in the descriptions are the amount of information. In other words, we are assuming the following:

**Assumption 6**: $\vee s_1 < t_1, s_2 < t_2 : f(s_1 + s_2, t_1 + t_2)$

$= \frac{t_1}{t_1+t_2} f(s_1,t_1) + \frac{t_2}{t_1+t_2} f(s_2,t_2)$

From the above assumptions,we can proved the following theorem:

**Similarity Theorem**: The similarity between $x$ and $y$ is measured by the ratio of the amount of information required to indicate the proximity between $x$ and $y$ and the information required to fully describe what $x$ and $y$ are:

$Sim(x,y) = \frac{logP(closeness(x,y))}{logP(description(x,y))}$

**Proof**: $f(s,t) = f(s+0, s+(t-s))$

$= \frac{s}{t} \times f(s,s) + \frac{t-s}{t} \times f(0,t-s),$ Assumption 6

$= \frac{s}{t} \times 1 + \frac{t-s}{t} \times 0 = \frac{s}{t},$ Assumption 4 and 5

Since similarity is the ratio between the amount of information in the proximity and the amount of information in the description of two objects, their similarity tells us how much more information is needed to determine what these two objects are. While most measures of similarity increase with closeness and decrease with difference,similarity of distance only decreases with increase in difference.

100