Available online at www.isas.org.in/jisas

JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 64(2) 2010 289-293

On Some Aspects of Regression in Social Network Models

Bikas K. Sinha

Applied Statistics Division, Indian Statistical Institute, Kolkata

Received 09 February 2010; Revised 01 April 2010; Accepted 13 April 2010

SUMMARY

This paper deals with some probabilistic aspects of social networks, viewed as directed graphs or digraphs. The vertices are connected by one-way or two-way edges. A simple model for random generation of such edges between pairs of vertices is discussed. The concepts of out-degrees and in-degrees of vertices are mentioned. The total of out-degree and in-degree of a vertex may be regarded as the extent of "total number of moves" experienced by that vertex. Distributional aspects of these vertex-oriented random variables are discussed with special reference to the nature of regression.

Keywords: Regression, Digraph, Directed graph, Out-degree, In-degree, Density.

1. INTRODUCTION

Social network models are well known to some extent in the community of quantitative sociologists. Though there is considerable interplay of sociology, graph theory and statistics towards our understanding of social network analysis '(sna)', by and large graph-theorists and statisticians have not been that much involved in the research component of sna. The purpose of this paper is to take up one simple statistical model available in the literature and to present some aspects of regression hitherto not explicitly available in the published literature.

A social network, viewed as a "digraph" or a "directed graph", deals with a given number of "units" which are regarded as "vertices" in the graph-theoretic terminology. Between any two given units i and j, there could be an "edge" originating at i and terminating at j. This edge has a "direction" viz., from i to j in which case, we say that there is a one-way "tie" from i to j. In case there is also an edge in the "reverse" direction, that is, originating at j and terminating at i, we say that there is a "reciprocal" relation between the two vertices i and j. In other words, there is "reciprocity" between the pair of units i and j. In general, between a given pair of vertices i and j, there could be no edge at all, or either way only one edge and not the other [i.e,, only one one-way tie], or else, a two-way tie i.e., a reciprocal relation. The vertices, if at all, at the "receiving end" of the ties originating at unit i are said to constitute the "Out-Degree Set" of unit i and the number of such vertices is called the "out-degree" of the unit i, denoted by d_i. Clearly, d_i can assume the values 0, 1, 2, ..., $N-1$ where N denotes the total number of units under consideration. Similarly, "In-Degree Set" of unit i refers to the set of vertices, if any, which serve as "originating" vertices of edges with i as the "terminating" vertex. We denote by e_i the "indegree" of unit i which refers to the cardinality of the In-Degree Set referred to above. It follows that once again, e_i assumes the values 0, 1, 2, ..., $N-1$ and further that $\sum d_i = \sum e_i = m$, say. Clearly, m can be interpreted as the total number of ties or edges in the network, with a "double-count" for every reciprocal tie. An interesting situation arises when for a given unit i, it so happens

E-mail address : sinhabikas@yahoo.com

that $d_i = e_i = 0$. That means, unit i does *not* interact with any other unit in the population under consideration in either direction. Such units are called "isolates". Naturally, isolates form a special class by themselves and sociologists need to provide critical and reasonable explanations towards formation of such classes.

It is argued that two most basic parameters of a social network are the *number of vertices N* and the *total number of edges m*. The ratio $m/N(N-1)$ is referred to as the "density" of the relationship among the vertices in the net-work. It is also regarded as the "global density" of a network. Further to these, we have the sets of out-degrees and in-degrees of the vertices. As discussed before, when there is a reciprocal tie, the edges in the pair of vertices enter into a "symmetrical" relation. The total number of such symmetric or reciprocal ties is also a parameter of interest to the sociologists. We will concentrate on the out-degree and in-degree parameters of a social network and study some aspects of regression, after formulating the models for generating such networks.

Since the edges in the network [also referred to as "arcs" in the network] may not be distributed uniformly over the vertices, one may be interested in the corresponding vertex-specific "local" measure which we call the *local density*. What is the counterpart of m for the i-th vertex? It is the number of ties d_i going out from it i.e., its out-degree. How do we standardize d_i? Since the minimum and maximum values d_i can take are 0 and $N-1$, we take $d_i/(N-1)$ to be the local density of the i-th vertex.

There is another possible counterpart of m for the i-th vertex: the number of ties e_i coming to it, i.e., its in-degree. The corresponding local measure of density is $e_i/(N-1)$. We may note that d_i and e_i signify entirely different matters. In the words of sociologists, out-degree represents "expansiveness" of the units while indegree signifies their "popularity". However, it is easy to see that the sum of the d_i's, as well as the sum of the e_i's, is equal to m. Hence the global density is the average of the local densities of the vertices, whether they are based on the out-degrees or on the in-degrees. Yet, one traditionally takes $d_i/(N-1)$, rather than $e_i/(N-1)$ as the local measure, particularly when the network represents a sociological choice relation.

An excellent expository paper is due to Rao and Bandyopadhyay (1987) wherein all the above concepts are explained with illustrative examples. For large networks, it may be prohibitive to collect data on all pairwise network patterns. Hence sampling from a network becomes essential for a global understanding and for estimation of meaningful parameters therefrom. There is an impressive literature in this direction. A few references are listed at the end. [Frank (1977a, 1977b, 1978), Feller (1960), Goswami *et al*. (1990), Sinha (1997), Thompson (2006)]. Some combinatorial problems associated with constructions of networks from given sets of parameters are challenging and research has flourished in that direction as well. See, for example, Achuthan *et al*. (1984). A text book on sna is also worth citing: Wassermann and Faust (1994). A forthcoming book is by Bandyopadhyay *et al*. (2010).

2. SOCIAL NETWORK MODEL OF INDEPENDENCE

Over the years, several graph-theoretic and probabilistic models for understanding social networks have evolved. A comprehensive review can be found in Bandyopadhyay *et al*. (2010).

Herein we will discuss only one stochastic or probabilistic model and we will relate it to the study of regression.

This simplest stochastic model assumes that there is a 50-50 chance of an edge-formation from unit i to unit j and that in this process all ordered pairs behave independently of one another. Thus, the chance of a reciprocal tie involving a pair of vertices is 0.25. Clearly, (i) the distribution of m is Binomial with parameters $(N(N-1); 0.5)$; (ii) the distribution of out-degree [same as in-degree] for any unit is again Binomial with parameters $(N-1; 0.5)$; and (iii) the distribution of s is also Binomial with parameters $(N(N-1)/2, 0.25)$.

Generalizing this model, it is assumed that there is a chance of p for formation of an edge from unit i to unit j and that all "edge-formations" between pairs of units are independent. Thus, m will be distributed as Binomial with parameters $(N(N-1), p)$; d_i [same as e_i] will be distributed as Binomial with parameters $(N-1, p)$; s will be distributed as Binomial with parameters $(N(N-1)/2, p^2)$ and so on.

Recall that d_i is the number of edges originating from unit i while e_i is the number of edges terminating at unit i. Therefore, $D_i = d_i + e_i$ represents the "total number of moves" experienced by the i-th unit. In other words, the i-th unit is specifically involved in a subtotal of D_i ties in the network. We will be particularly interested in the joint distribution of the D_i's.

According to the model, each of d_i and e_i is Binomial with parameters $(N-1, p)$ and the two are independent. Therefore, each D_i is distributed as Binomial with parameters $(2(N-1), p)$. Consider a pair of D's viz., D_i and D_j for $i \neq j$. Note that of the four constituents of D_i and D_j, (i) d_i and d_j are independent; (ii) e_i and e_j are independent; (iii) d_i and e_i are independent; (iv) d_j and e_j are independent. However, d_i and e_j are *not* statistically independent; so is the other pair viz., d_j and e_i.

Given these facts, we propose to investigate the nature of the joint distribution of D_i and D_j and examine therefrom the nature of the regression of D_i on D_j.

3. NATURE OF REGRESSION IN A SOCIAL NETWORK MODEL OF INDEPENDENCE

Once the vertices i and j are fixed, there are four possibilities as to their mutual status with regard to out-degree and in-degree:

Case I : Units i and j are detached from each other;

Case II : There is a tie originating at i and terminating at j but there is no reciprocal tie;

Case III : Same as Case II with the roles of i and j reversed;

Case IV : There are two ties mutually connecting both i and j.

The probabilities of these cases are given by q^2; pq; pq and p^2 respectively.

Note that the results are independent of the specific choice of the two vertices.

We are now in a position to deduce the form of the joint distribution of D_i and D_j under each of the above cases as follows :

(I) Under Case I :

$$P_I[r, s] = P[D_i = r, D_j = s]$$

$$= \binom{(2N-4)}{r} \cdot \binom{(2N-4)}{s} \cdot p^{(r+s)} q^{(4N-8-r-s)} \cdot q^2$$

$$= \binom{(2N-4)}{r} \cdot \binom{(2N-4)}{s} \cdot p^{(r+s)} q^{(4N-6-r-s)}$$

(II) Under case II:

$$P_{II}[r, s] = P[D_i = r, D_j = s]$$

$$= \binom{(2N-4)}{(r-1)} \cdot \binom{(2N-4)}{(s-1)} \cdot p^{(r+s-2)} q^{(4N-6-r-s)} \cdot qp$$

$$= \binom{(2N-4)}{(r-1)} \cdot \binom{(2N-4)}{(s-1)} \cdot p^{(r+s-1)} q^{(4N-5-r-s)}$$

(III) Under Case III: Same as in Case II

(IV) Under Case IV:

$$P_{IV}[r, s] = P[D_i = r, D_j = s]$$

$$= \binom{(2N-4)}{(r-2)} \cdot \binom{(2N-4)}{(s-2)} \cdot p^{(r+s-4)} q^{(4N-4-r-s)} \cdot p^2$$

$$= \binom{(2N-4)}{(r-2)} \cdot \binom{(2N-4)}{(s-2)} \cdot p^{(r+s-2)} q^{(4N-4-r-s)}$$

Therefore, $P[r, s]$ is now computed as the sum of the above four expressions and we obtain

$$P[r, s] = \binom{(2N-4)}{r} \cdot \binom{(2N-4)}{s} \cdot p^{(r+s)} q^{(4N-6-r-s)} + 2\binom{(2N-4)}{(r-1)} \cdot \binom{(2N-4)}{(s-1)} \cdot p^{(r+s-1)} q^{(4N-5-r-s)} + \binom{(2N-4)}{(r-2)} \cdot \binom{(2N-4)}{(s-2)} \cdot p^{(r+s-2)} q^{(4N-4-r-s)}$$

From this, we derive $P[s, .] = \binom{(2N-2)}{s} p^s q^{(2N-s)}$ $= P[.,s]$, as expected.

Hence, $P[r \mid s] = P[r, s]/P[., s]$ defines the conditional distribution of one of the two D_i's, given the other is equal to s.

This yields:

$$E[r \mid s] = \left[\binom{(2N-4)}{s} / \binom{(2N-2)}{s}\right] \cdot \sum_r \left[r \cdot \binom{(2N-4)}{r} p^{(r)} q^{(2N-4-r)}\right] + 2\left[\binom{(2N-4)}{(s-1)} / \binom{(2N-2)}{s}\right] \sum_r \left[r \cdot \binom{(2N-4)}{(r-1)} p^{(r-1)} q^{(2N-3-r)}\right]$$

$$+\frac{\binom{(2N-4)}{(s-2)}}{\binom{(2N-2)}{s}}\cdot\sum_r[r\cdot\binom{(2N-4)}{(r-2)}p^{(r-2)}q^{(2N-2-r)}]$$

$= [(2N-2-s)(2N-3-s)/(2N-2)(2N-3)]$
$(2N-4)p + 2.[s(2N-2-s)/(2N-2)(2N-3)]$
$[(2N-4)p+1] + [s(s-1)/(2N-2)(2N-3)]$
$[(2N-4)p+2]$

$= (2N-4)p + 2s/(2N-2)$, upon simplification.

Hence the regression of D_i on D_j is linear in $D_j = s$.

We readily verify that

$E[r] = EE[r \mid s] = (2N-4)p + 2(2N-2)p/(2N-2) = (2N-2)p$, as expected.

We now specialize to the particular case of $p = 0.5$ and discuss some further results along the line of regression.

First of all, we display below the joint distribution of D_i, D_j, D_k for $i \neq j \neq k$.

Let $A[r, s, t] = [\binom{(2N-6)}{r}][\binom{(2N-6)}{s}][\binom{(2N-6)}{t}]$.

Then $2^{(6N-12)}P[r, s, t] = [A(r, s, t) + 2A(r-1, s-1, t) + A(r-1, s, t-1) + A(r, s-1, t-1) + A(r-2, s-2, t) + A(r-2, s, t-2) + A(r, s-2, t-2)$

$+ 4A(r-2, s-1, t-1) + A(r-1, s-2, t-1)$
$+ A(r-1, s-1, t-2) + 2A(r-3, s-2, t-1)$
$+ A(r-3, s-1, t-2) + A(r-2, s-3, t-1)$
$+ A(r-2, s-1, t-3) + A(r-1, s-3, t-2)$
$+ A(r-1, s-2, t-3) + A(r-4, s-2, t-2)$
$+ A(r-2, s-4, t-2) + A(r-2, s-2, t-4)$
$+ 8A(r-2, s-2, t-2) + 4A(r-3, s-3, t-2)$
$+ A(r-3, s-2, t-3) + A(r-2, s-3, t-3)$
$+ 2A(r-4, s-3, t-3) + A(r-3, s-4, t-3)$
$+ A(r-3, s-3, t-4) + A(r-4, s-4, t-4)]$

There are altogether 64 terms, not all distinct.

From the above, the bivariate marginals are derived as

(1) $P[0, 0] = (0.5)^{2(2N-3)} > P[0]P[0]$;

(2) $P[1, 0] = (0.5)^{2(2N-3)} \cdot [\binom{(2N-4)}{1}] > P[1]\cdot P[0]$;

(3) $P[1, s] = [(0.5)^{(4N-7)}\ [\binom{(2N-4)}{(s-1)}] + [(0.5)^{(4N-6)}$
$[\binom{(2N-4)}{s}][\binom{(2N-4)}{1}]$; $s \geq 1$;

(4) $P[r, s] = [(0.5)^{(4N-6)}\ [\binom{(2N-4)}{r}][\binom{(2N-4)}{s}]$
$+ [(0.5)^{(4N-7)}\ [\binom{(2N-4)}{(r-1)}][\binom{(2N-4)}{(s-1)}]$
$+ [(0.5)^{(4N-6)}\ [\binom{(2N-4)}{(r-2)}][\binom{(2N-4)}{(s-2)}]$; $r, s \geq 2$.

Remark 1. $P[r, s]$ can also be derived from $P[r, s, t]$ given above.

Here are the univariate marginals:

$P[r] = \sum_s P[r, s] = (0.5)^{(2N-2)}.\binom{(2N-2)}{r}$, upon simplification.

It follows that $P[0, 0] > P[0]P[0]$; However, for $s = 2N-3$, for example, the inequality goes in the other direction!

4. SUFFICIENT STATISTICS IN THE SOCIAL NETWORK MODEL OF INDEPENDENCE

Recall $D_r = d_r + e_r$, $r = 1, 2, \ldots, N$. Set $D = [D_1, D_2, \ldots, D_N]$.

We now make the following claim: D is a sufficient statistic for p in the model of independence.

Proof. First note that the likelihood for a Graph G under the assumed model is given by the product of terms of the type

$$L_{(r, s)}[p \mid G] = p^{I(r, s)}(1-p)^{(1-I(r,s))}$$

where

$I(r, s) = 1$ if r goes to s
$= 0$ o.w.

This leads to the likelihood ratio as

$$\theta\sum\sum_{(r,s)}{}^{I(r, s)}(1-p)^{N(N-1)}$$

where $\theta = p/(1-p)$

Next note that

$$\sum\sum_{(r,s)}I(r, s) = \sum_r d_r = \sum_r e_r = \sum_r D_r/2$$

Hence, d_r's, e_r's as also D_r's are sufficient statistics.

Given d_r'*s* we can work out the conditional distribution of the e_r'*s* and hence of the D_r'*s*. Likewise, from e_r's, we can work out conditional distributions of d_r'*s* and of the D_r'*s*. These seem to be relatively straightforward ! However, from D_r's, it seems difficult to find conditional distributions of d_r's [and hence of the e_r's]. It is, of course, trivial to note that for a fixed r, the conditional distribution of d_r, given D_r, is Hypergeometric and the same is true for e_r. In fact, both the conditional distributions are identical. For given vector of D_r's, the joint conditional distribution of the vector of d_r's is independent of the parameter p, but the exact form of this distribution is complicated.

Here is an illustrative example to show the level of computational complexity. Consider $N = 3$ and assume the D_r-values to be [2, 2, 2]. Then possible values of d_r's are :

(i) [2, 1, 0]; (ii) [0, 1, 2]; (iii) [1, 1, 1]

and the corresponding incidence matrices are formed as :

(i) $I(1, 2) = 1$; $I(1, 3) = 1$; $I(2, 3) = 1$; rest all 0's

(ii) $I(2, 1) = 1$; $I(3, 1) = 1 = I(3, 2)$; rest all 0's

(iiia) $I(1, 3) = 1$; $I(2, 1) = 1$; $I(3, 2) = 1$; rest all 0's

(iiib) $I(1, 2) = 1$; $I(2, 3) = 1$; $I(3, 1) = 1$; rest all 0's

It turns out that the conditional distributions of d_r's and e_r's are the same and it is given by

$P[2, 1, 0 \mid D] = P[0, 1, 2 \mid D] = 1/4$
$P[1, 1, 1 \mid D] = 1/2$

ACKNOWLEDGEMENT

The author is highly thankful to the anonymous referee for constructive suggestions on an earlier version of the manuscript.

REFERENCES

Achuthan, N., Rao, S.B. and Rao, A.R. (1984). The number of symmetric edges in a digraph with prescribed out-degrees. *Comb. Appl.*, Proc. of the Seminar in honour of Professor S.S. Shrikhande, Indian Statistical Institute, Calcutta, 8-20.

Bandyopadhyay, S., Rao, A.R. and Sinha, Bikas K. (2010). *Social Network Models with Statistical Applications*. Sage Publications.

Frank, O. (1977a). Survey sampling in graphs. *J. Statist. Plann. Inf.*, **1**, 235-264.

Frank, O. (1977b). Estimation of graph totals. *Scandinavian J. Statist.*, **4**, 81-89.

Frank, O. (1978). Sampling and estimation in large social networks. *Social Networks*, **1**, 91-101.

Feller, W. (1960). *An Introduction to Probability Theory and its Applications*. Vol. 1, Asia Publ. House.

Goswami, A., Sengupta, S. and Sinha, Bikas K. (1990). Optimal strategies in sampling from a social network. *Sequential Anal.*, **9**, 1-18.

Rao, A. Ramachandra and Bandyopadhyay, Suraj (1987). Measures of reciprocity in a social network. *Sankhyā*, **A49**, 141-188.

Sinha, B.K. (1997). Some inference aspects of a social network. In: *Applied Statistical Science, II*. Malang, 77-86, Nova Sci. Publ., Commack, New York.

Thompson, S.K. (2006). Targeted random walk designs. *Survey Methodology*, **32**, 11-24.

Wassermann, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1-24.