International Journal of STD & AIDS
http://std.sagepub.com/

Combining social network analysis and cluster analysis to identify sexual network types

E De Rubeis, J L Wylie, D W Cameron, R C Nair and A M Jolly

Int J STD AIDS 2007 18: 754
DOI: 10.1258/095646207782212234

ORIGINAL RESEARCH ARTICLE

Combining social network analysis and cluster analysis to identify sexual network types

E De Rubeis MSc*, **J L Wylie** PhD†‡§, **D W Cameron** MD**, **R C Nair** PhD* **and A M Jolly** PhD*††

*Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, ON; †Cadham Provincial Laboratory, Manitoba Health; ‡Department of Medical Microbiology; §Department of Community Health Sciences, University of Manitoba, Winnipeg, MB; **The University of Ottawa at the Ottawa Hospital, Ottawa, ON; ††Division of Modeling and Projections, Centre for Infectious Disease Prevention and Control, Public Health Agency of Canada, Ottawa, ON, Canada

Summary: Increases in the rates of sexually transmitted infections (STIs) suggest that control programmes may not be effectively targeting diverse subpopulations. The objective of this investigation was to examine STI transmission within different groups, using both social network analysis and cluster analysis. Routine partner notification data were analysed from individuals diagnosed with, or exposed to an STI in Manitoba. Groups were identified and characterized. Three different clusters of groups were identified, comprised of demographically and clinically distinct individuals. A greater understanding of disease transmission patterns within these groups will aid in the development of targeted education and prevention programmes for all STIs.

Keywords: sexually transmitted diseases, epidemiology, cluster analysis, contract tracing

INTRODUCTION

Chlamydia trachomatis and *Neisseria gonorrhoeae* are responsible for significant morbidity. In attempts to decrease the burden of disease, population-based control programmes were initiated throughout Canada. However, approximately over the last 10 years rates of both infections have begun increasing, and several provinces have experienced outbreaks of locally acquired infectious syphilis.

Social network analysis (SNA) has been used as a tool in understanding the transmission of sexually transmitted infections (STIs).[1-8] The basis of SNA is that 'the structure of a network has consequences for its individual members and for the network as a whole over and above effects of characteristics and behaviours of the individuals involved'.[2] A network consists of a set of nodes representing people, connected by a set of edges representing relationships.[9,10] A network can be partitioned into components, defined as a group of individuals within the network, connected either directly or indirectly through sex.[11]

In 1997, Jolly and Wylie used SNA to study transmission patterns of STIs in the province of Manitoba using routinely collected surveillance data.[12–16] They identified two distinct component types, linear and radial. Linear components were characterized by a low variation in the number of sexual partners within a component, with no central individual, and radial components were characterized by one central individual connected to many immediate individuals with one or two partners. More significantly, linear components had a higher proportion of aboriginal members, higher positivity rates and the majority contained both chlamydia and gonorrhoea.

This indicated that distinct sexual networks exist, potentially requiring different programmatic approaches for disease control. Here we used cluster analysis (CA) to formally characterize the differences among components, grouping together components with similar profiles. Approaches of this type can offer a better understanding of the differences among distinct subpopulations and provide information relevant to the development of targeted, population-specific prevention programmes.

METHODS

Data source

Routinely collected STI surveillance data were obtained from the Communicable Disease Control Unit of Manitoba Health. Collection of this data through partner notification techniques has been previously described in detail.[12–16] Data on laboratory-confirmed cases from June 2002 to

Correspondence to: Mrs Emily De Rubeis, 451 Smyth Road, Ottawa, ON, Canada K1H 8M5
Email: emily.derubeis@uottawa.ca

October 2003 and nominated sexual partners from October 2002 to October 2003 exposed to an STI were extracted. Age, gender, region of residence, infection status and ethnicity were extracted. Only aboriginal status was available for ethnicity; therefore, this variable consists of aboriginal versus non-aboriginal ethnicity.

As a person could have multiple roles, for example, a contact could become a case, the data were de-duplicated so the unit of observation was a person and not an event; unique identifiers were assigned and identifying information was removed.[16]

Statistical analysis

The data were imported into SAS, and cleaning and validating were completed.[17] Data cleaning consisted of examining descriptive statistics, quantifying the amount of missing data and identifying outliers. For both cases and contacts, missing data for place of residence (0.3% and 12%, respectively) were imputed using an individual's designated Regional Health Authority. Ages of 48% of contacts were missing; no changes were made to these records.

A visual basic program (©Ann Jolly 2003) was used to format the data into a PAJEK input file in which the components were identified. PAJEK is a social network computer program specifically designed to handle large data-sets.

Collective responses to sociodemographic, clinical and geographic variables were calculated for each component, defining, for example the proportion of members within a component who were identified as aboriginal.

Cluster analysis

All steps of the CA were completed in SAS. Clustering units ($n = 239$) were components containing five or more people, excluding six components that contained an anonymous participant. The size restriction was placed in attempts to minimize ties during the clustering procedure, which occurs when components have identical profiles.

Three clustering variables were used: proportion of component members who were aboriginal; mean degree centralization and proportion of individuals residing in the same geographic area.[16] Degree centralization, a measure of the variance in distribution of degree centrality within a component, was used to objectively distinguish between the linear and radial components previously described by Wylie and Jolly.[16] Geographic residence area is related to STI transmission via spatial bridging opportunities for STI. Ethnicity was included as a clustering variable, given its known correlation with network types.[16]

The approximate covariance estimation for clustering (ACECLUS) procedure, based on Mahalanobis' generalized distance, was used to measure the distance between pairs of components. This is an iterative process that uses pairwise differences to estimate the pooled within-cluster covariance matrix.[18,19] Prior to the initiation of this procedure, several criteria needed to be specified. The initial estimate for the pooled within-cluster covariance matrix was the total-sample covariance matrix obtained from the component by variable input matrix (239 components by 3 clustering variables). The distance cut-off (μ) was obtained by specifying the proportion of pairwise differences between observations that would be less than μ, and transforming this proportion (ρ) into the appropriate distance value. Three values of the proportion of pairwise differences, 0.2, 0.10 and 0.05, were selected. As suggested in the literature, a value of 0.001 or smaller was used to indicate sufficient closeness between successive estimates of the pooled within-cluster covariance matrix.[19] The resulting standardized canonical coefficients were used to linearly transform the raw data. The three resulting data-sets, one for each specification of ρ, were used as input matrices for clustering.

CA was completed using Ward's[20] Minimum Variance method, forming clusters on the basis of loss of information. The algorithm begins with each component representing a single cluster; a situation in which the largest amount of information is available. Next, the algorithm considers all possible clusters that could be formed, joining the two that result in a minimal loss of information. This algorithm is repeated until the data are systematically reduced to one cluster containing all 239 components.

The optimal clustering level, restricted to the final 10 clustering levels, was determined by examining both dendrograms and two numeric stopping rules (Pseudo F and T^2 tests).[21,22] The suggested stopping points for the Pseudo F and T^2 tests were absolute or local maxima and local minima, respectively. We then examined the characteristics of the clusters using variables external to the solution.

Ethical approval

Ethical approval was granted by the Health Research Ethics Board, University of Manitoba, participating Regional Health Authorities in Manitoba, the Health Information Privacy Committee of Manitoba Health and the Ottawa Hospital Research Ethics Board.

RESULTS

Description of the sexual network

The entire study population consisted of 8476 uniquely identified individuals with 4683 cases and 3793 contacts (Table 1). Nearly one-third (33%) of the cases were men and 29% were aboriginal. Over 70% of participants had chlamydia; men were more likely to be diagnosed with gonorrhoea (19%) than women (7%). Over the course of the study, nearly one-tenth of men and women were diagnosed with a second STI. In total, 70% of the contacts were men. PAJEK identified 2508 components of size two or greater, the largest containing 33 individuals (Table 2). In total, 60% of the components contained two members; 23% contained three members and 35 components of size 10 or greater were identified. Additionally, 1192 laboratory-confirmed cases named no sexual partners.

Table 1 **Demographic and clinical characteristics of cases and contacts in the identified sexual network (June 2002 to September 2003)**

	Cases (*n*=4683)*		**Contacts (*n*=3793)***	
	Men (*n*=1525)	**Women (*n*=3123)**	**Men (*n*=2658)**	**Women (*n*=981)**
Characteristics	***n* (%)**			
Age (years)	Median=23 (IQR=19, 28)†	Median=20 (IQR=18, 24)	Median=23 (IQR=20, 28)	Median=21 (IQR=18, 27)
Residence				
Winnipeg	827 (54)	1565 (50)	1189 (45)	400 (41)
Southern Manitoba‡	286 (19)	592 (19)	468 (18)	149 (15)
Northern Manitoba‡	398 (26)	938 (30)	672 (25)	242 (25)
Outside Manitoba	13 (1)	27 (1)	172 (6)	78 (8)
First Nations	439 (29)	916 (29)	318 (12)	72 (7)
Infection acquired				
Chlamydia	1064 (70)	2636 (84)	–	–
Gonorrhoea	285 (19)	220 (7)	NA	NA
Syphilis	5 (1)	6 (1)	–	–
Chlamydia and gonorrhoea	168 (11)	253 (8)	–	–
Repeat case	100 (7)	246 (8)	NA	NA
Co-infection	152 (10)	221 (7)	NA	NA
Repeat contact	96 (6)	55 (2)	107 (4)	15 (1)

*In total, 35 cases and 154 contacts with unknown gender were not included; participants with missing/incomplete data were not included

†IQR=interquartile range

‡Southern Manitoba includes Brandon, South Eastman, Interlake, Central, Assiniboine and Parkland Regional Health Authorities. Northern Manitoba includes North Eastman, Norman, Burntwood and Churchill Regional Health Authorities

Table 2 **Frequency distribution of *n*=2508 network components identified in Manitoba (June 2002 to September 2003)**

Component size	**Frequency (%)**	**Cumulative percent of individuals**
2	1512 (60)	3024 (41)
3	577 (23)	4755 (65)
4	174 (7)	5451 (75)
5	89 (4)	5896 (81)
6	53 (2)	6214 (85)
7	34 (1)	6452 (89)
8	20 (1)	6612 (91)
9	14 (1)	6738 (93)
⩾10	35 (1)	7284 (100)

Cluster analysis

Dendrograms and numeric stopping rules were used to determine the optimal clustering level. When the distance value (p) was set at 0.20, there were no local maxima for the Pseudo F test (Table 3); with p at 0.10, two local maxima for the Pseudo F existed at levels 5 and 3; whereas the Pseudo T^2 test indicated optimal clustering at an additional three levels; with p at 0.05, local maxima were observed but the pooled within-cluster covariance matrix was unstable. Examination of the resultant dendrogram ($p = 0.10$) revealed large differences between consecutive R-squared values, representing the proportion of variance attributed to clustering, which occurred at the last three clustering levels (Figure 1). In view of the consensus between the stopping rules when p was set at 0.10 and the dendrogram, the optimal number of clusters was determined to be three.

The three clusters contained 123, 69 and 47 components, respectively (Table 4). For the three clustering variables, the proportion of aboriginal people increased from cluster 1 to cluster 3, as did the percentage of individuals within a component who lived in the same geographic region. Mean degree centralization decreased across the three clusters.

For the variables external to the clustering solution, notable differences were seen for cluster 3 which contained the widest mean age range (10 years), the highest mean percentage of individuals living in northern Manitoba and the highest mean percentage of repeatedly named contacts. Cluster 2 contained the highest mean percentage of components with gonorrhoea alone (15%) and included all syphilis cases (data not shown). Cluster 1 contained components with the largest mean proportion of same-sex partnerships.

DISCUSSION

Recently, Canada has witnessed an upsurge in the incidence of bacterial STIs. Although a proportion of the increase may be attributed to the growing use of less-invasive testing procedures for chlamydia and gonorrhoea, with greater diagnostic sensitivity, the increase in incidence of these STIs as a group and the re-emergence of syphilis indicate that there are other contributing factors.

It has been hypothesized that over time, within a population in which control programmes have been initiated, the reservoir of infection that supports an epidemic becomes concentrated within marginalized, hard-to-reach subpopulations who have limited contact with the health-care system.[23] Hence, there is a need for

Table 3 **Pseudo *F* and T^2 stopping rules obtained for the final 10 clustering levels using Ward's Minimum Variance Clustering method with three different input matrices corresponding to the different values of *p* used during the approximate covariance estimation for clustering (ACECLUS) procedure**

	***p*=0.2**		***p*=0.1**		***p*=0.05**	
Clustering level	**Pseudo *F***	**Pseudo T^2**	**Pseudo *F***	**Pseudo T^2**	**Pseudo *F***	**Pseudo T^2**
10	187	33	485	28	677	17
9	191	14	506	19	698	22
8	191	31	518	25	710	35
7	191	29	520	31	716	71
6	196	75	546	63	706	54
5	210	27	564	66	716	30
4	238	49	535	89	732	100
3	240	95	595	67	489	132
2	319	66	540	259	543	423
1	–	319	–	540	–	543

Clustering level 1 refers to the point at which all *n*=239 components are grouped into one cluster

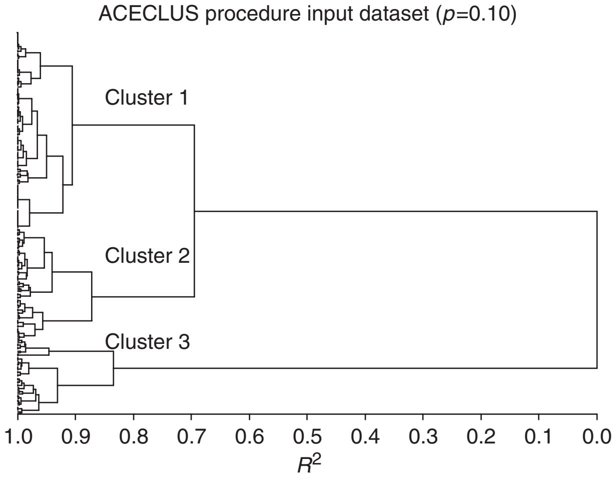

Figure 1 **Dendrogram using input data-set variables from the approximate covariance estimation for clustering (ACECLUS) procedure, where $p=0.1$. The algorithm proceeds from left to right, $n=239$ clusters to $n=1$ cluster, respectively; each vertical line corresponds to the joining of two clusters and the respective R^2 is indicated on the horizontal axis**

prevention strategies that target these subpopulations and have an impact beyond these groups by reducing secondary transmission.[24] However, targeted prevention strategies cannot be implemented without proper characterization of the subpopulations present within an area. The current study begins exploring novel ways to characterize the broad subgroups that exist within a larger sexual network by combining SNA with CA. To the authors' knowledge, it is the first to use this combined approach.

CA identified three distinct clusters that differed on the basis of geography, the proportion of laboratory-confirmed cases, infecting pathogen and structure. There were distinct differences in the geography of the clusters – components within Cluster 1 contained individuals living primarily in Winnipeg; components within Cluster 2 contained individuals geographically dispersed throughout the entire province; whereas most of the individuals within Cluster 3 lived in northern Manitoba. These findings indicate that the endemic transmission of STIs in Winnipeg and northern Manitoba may be driven largely by the individuals within Clusters 1 and 3, whereas individuals within Cluster 2 components may be acting as geographic bridges for STIs between the various areas of the province.

The mean proportion of cases differed among the three clusters. Components in Cluster 1 had the lowest mean proportion of cases, despite the moderately high proportion of individuals with repeat infections and those repeatedly nominated. These components were often star-like or radial. This pattern suggests that partner notification and follow-up may be difficult with this group such that some cases become repeatedly infected from unidentified or untested contacts. However, the relatively low percentage of cases also suggests that behavioural patterns in this group limit transmission to many of the individuals within this group. Components within Cluster 3 had the highest proportion of cases and repeatedly nominated members. Despite the problems with STI follow-up in the north, the high number of diagnosed cases suggests either effective case finding or that the structure of these networks is particularly conducive to STI spread.

Interestingly, clusters differed with respect to the infecting pathogen. This pattern is important as it suggests that different behaviours within these clusters may facilitate or hinder the spread of some STI pathogens. Components within Cluster 2 had a high proportion of gonorrhoea cases, and contained all of the syphilis cases within our data-set. This potential pathogen-specificity in terms of transmission is an area that would be of particular importance for further research.

Differences in the degree centralization among the different clusters suggest different control programmes may be advantageous in these distinct populations. Programmes aimed at the radial components within Cluster 1 may be most effective if directed towards central individuals, whereas more general programmes aimed at all individuals may be needed for components of the type seen in Cluster 3, where the majority of components had linear structure.

In conclusion, these findings suggest that broadly distinct groups of individuals exist within the larger

Table 4 **comparison of the demographic, clinical and network characteristics of the three identified clusters**

	Cluster 1 (*n*=123)	**Cluster 2 (*n*=69)**	**Cluster 3 (*n*=47)**
Characteristics	**Mean (range)**	**Mean (range)**	**Mean (range)**
Clustering variables			
Mean % First Nations	0.7 (0, 11)	23 (11, 40)	64 (44, 100)
Mean degree centralization	0.7 (0.1, 1)	0.6 (0.1, 1)	0.5 (0.1, 1)
Mean % in same region	77 (14, 100)	80 (25, 100)	90 (44, 100)
Variables external to the clustering solution			
Mean age range (years)	7 (0, 40)	7 (0, 52)	10 (2, 39)
Mean % women	41 (0, 86)	46 (12, 87)	45 (17, 80)
Mean % unknown gender	3 (0, 86)	2 (0, 75)	0.4 (0, 20)
Mean % in Winnipeg	56 (0, 100)	36 (0, 100)	10 (0, 83)
Mean % in southern Manitoba	15 (0, 100)	22 (0, 100)	20 (0, 100)
Mean % in northern Manitoba	14 (0, 100)	35 (0, 100)	69 (0, 100)
Mean % outside Manitoba	6 (0, 86)	3 (0, 40)	1 (0, 20)
Mean % cases	36 (12, 80)	42 (12, 80)	51 (17, 80)
Mean % cases with chlamydia	83 (0, 100)	60 (0, 100)	69 (0, 100)
Mean % cases with gonorrhoea	7 (0, 100)	15 (0, 100)	9 (0, 50)
Mean % cases with CT and GC*	10 (0, 100)	20 (0, 100)	22 (0, 80)
Mean % repeat cases	15 (0, 100)	18 (0, 100)	16 (0, 100)
Mean % repeat contacts	10 (0, 67)	11 (0, 44)	21 (0, 67)
Mean component size	7 (5, 33)	8 (5, 32)	9 (5, 25)
Mean % same-sex partnerships	3 (0, 100)	1 (0, 25)	1 (0, 20)
Degree centralization	***n* (%)**	***n* (%)**	***n* (%)**
0–0.25	12 (9.8)	9 (13.0)	11 (23.4)
0.26–0.50	20 (16.3)	15 (21.7)	14 (29.8)
0.51–0.75	40 (32.5)	25 (36.2)	16 (34.0)
⩾0.76	51 (41.5)	20 (29.0)	6 (12.8)

*CT=*Chlamydia trachomatis*; GC=gonorrhoea

population that makes up the sexual network for an area. It should be cautioned that CA is an exploratory technique. It is intended to be hypothesis generating, rather than hypothesis confirming. As such, further research would be needed to examine in more detail the patterns identified above. In addition to the potential for pathogen-specific transmission within cluster types, as noted above, additional research of particular interest would focus on whether cluster-specific behaviours exist that drive the degree centralization patterns seen (which ultimately may translate into the differences in pathogen transmission). Research must also be directed at how conclusions obtained from broad social epidemiologic investigations of the type described above can be translated into methods or tools useful for public health workers when they are conducting partner notification activities at the level of the individual. In general, CA could be a useful tool in other health regions to better understand the broadly different types of sexual networks that exist and aid in the creation of network-specific control programmes. The same control programme may not be effective for all groups and targeted, network-specific education and prevention programmes may be needed for effective STI control.

ACKNOWLEDGEMENTS

The project was supported by grants from the Canadian Institutes of Health Research. DWC was supported by an Ontario HIV Treatment Network Career Scientist Award and the Ontario Ministry of Health.

REFERENCES

1 Friedman SR, Neaigus A, Jose B, *et al.* Sociometric risk networks and risk for HIV infection. *Am J Public Health* 1997;**87**:1289–96

2 Klovdahl AS. Social networks and the spread of infectious diseases: the AIDS example. *Soc Sci Med* 1985;**21**:1203–16

3 Klovdahl AS, Potterat JJ, Woodhouse DE, Muth JB, Muth SQ, Darrow WW. Social networks and infectious disease: the Colorado Springs Study. *Soc Sci Med* 1994;**38**:79–88

4 Laumann EO, Youm Y. Racial/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: a network explanation. *Sex Transm Dis* 1999;**26**:250–61

5 Neaigus A, Friedman SR, Jose B, *et al.* High-risk personal networks and syringe sharing risk factors for HIV infection among new drug injectors. *J Acquir Immune Defic Syndr Hum Retrovirol* 1996;**11**:499–509

6 Potterat JJ, Muth SQ, Bethea RP. Chronicle of a gang STD outbreak foretold. *Free Inq Creat Sociol* 1996;**24**:11–16

7 Rothenberg RB, Sterk C, Toomey KE, *et al.* Using social network and ethnographic tools to evaluate syphilis transmission. *Sex Transm Dis* 1998;**25**:154–60

8 Service SK, Blower SM. HIV transmission in sexual networks: an empirical analysis. *Proc R Soc Lond B Ser* 1995;**260**:237–44

9 Rothenberg RB, Potterat JJ, Woodhouse DE. Personal risk taking and the spread of disease: beyond core groups. *J Infect Dis* 1996;**174**(Suppl. 2): S144–9

10 Scott J. *Social Network Analysis: a Handbook*. 2nd edn. London: Sage, 2000

11 Potterat JJ, Rothenberg RB, Muth SQ. Network structural dynamics and infectious disease propagation. *Int J STD AIDS* 1999;**10**:182–5

12 Cabral T, Jolly AM, Wylie JL. *Chlamydia trachomatis omp1* genotypic diversity and concordance with sexual network data. *J Infect Dis* 2003;**187**:279–86

13 Jolly AM, Wylie JL. Sampling individuals with large sexual networks: an evaluation of four approaches. *Sex Transm Dis* 2001;**28**:200–7

14 Jolly AM, Wylie JL. Gonorrhoea and chlamydia core groups and sexual networks in Manitoba. *Sex Transm Infect* 2002;**78**(Suppl. 1):i145–51

15 Jolly AM, Muth SQ, Wylie JL, Potterat JJ. Sexual networks and sexually transmitted infections: a tale of two cities. *J Urban Health* 2001;**78**: 433–45

16 Wylie JL, Jolly AM. Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada. *Sex Transm Dis* 2001;**28**:14–24

17 SAS Institute Inc. *SAS V8*. Cary, NC: SAS Institute Inc, 1999

18 Art D, Gnanadesikan R, Kettenring JR. Data-based metrics for cluster analysis. *Utilitas Math* 1982;**21A**:75–99

19 SAS Institute Inc. *SAS OnlineDoc Version 8* [www.v8doc.sas.com/sashtml] 2003

20 Ward JH. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 1963;**58**:236–44

21 Calinski T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat* 1974;**3**:1–27

22 Duda RO, Hart PE. *Pattern Classification and Scene Analysis*. New York: Wiley, 1973

23 Wasserheit JN, Aral SO. The dynamic topology of sexually transmitted disease epidemics: implications for prevention strategies. *J Infect Dis* 1996;**174**(Suppl. 2):S201–13

24 Garnett GP, Bartley LM, Cameron DW, Anderson RM. Both a 'magic bullet' and good aim are required to link public health interests and health care needs in HIV infection. *Nat Med* 2000;**6**:261–2

(Accepted 9 May 2007)