Networks and Sex

11

The Use of Social Networks as Method and Substance in Researching Gay Men's Response to HIV/AIDS

Anthony P.M. Coxon

THIS CHAPTER IS BASED UPON THE VIEW that much behavioral research on gay men and their reaction to the AIDS pandemic is not so much psychologistic as overly individualistic in approach and that this severely limits its utility and prevents us from tackling the most challenging substantive and methodological (and hence policy-relevant) issues.[1]

This bias in the design of such studies often begins with the words: "Since a random sample of gay men is not possible, we used snowball sampling . . ."

They usually didn't—and the bias continues into the heartland of sexual behavior. Despite the building up of wave upon wave of prevalence figures for this and that sexual act, we still know little more than Kinsey did about the context of relationships.

A critical deficiency is the lack of a network perspective—network concepts, notions and methodologies. Now it may well be that networks are believed by the research community to be too complex to study and that they do not really add value sufficiently to compensate for their undoubted problems. In part this is true: it is often difficult enough to persuade an individual to participate in sexual behavior studies, without having to go on to recruit couples or—more relevantly—casual partners. It may also be that sex researchers shy away from the greater involvement and commitment of such detailed research methods, and prefer to rely upon the tried-and-tested survey.

I want to explore the applicability of network notions in three areas that loom large in our own research and in many other projects on sexual behavior. Network notions impinge most directly in the areas of:

- design,
- in networks of sexual contacts, and
- in identifying (unknown) sexual partners.

In so doing I shall be frank and realistic, so that others may learn from our mistakes and be a little more adventurous than funding authorities might sometimes like us to be!

1. Definitions and Sample Selection

From the outset, Project SIGMA resisted any attempt to produce a random sample of gay and bisexual men.[2] This was done on both principled and practical grounds. In principle, the tendency for policy-makers and others to think of "homosexuality" as a lasting and recognizable attribute was not to be encouraged, and Kinsey's research (1948:650–657) illustrates well how prevalence estimates of "homosexual men" can be made to range from four percent to almost fifty percent by successively relaxing the criteria of the type of sexual contact and the time-period of sexual involvement with those of the same sex (Coxon 1987).

But practical issues were paramount; no general population survey of sexual behavior was then envisaged in the U.K. and the cost of attempting to sample randomly on a two-stage basis (initially "combing" to produce a population frame, and secondly sampling within it) would be well outside funding agencies' means. In any event, the relevant population is narrower than the notional category of "homosexual men," being concerned with potential or actual HIV transmission rather than sexual identity. By this time, the causative agent of AIDS was known to be viral and the first crucial network-contact study (Auerbach 1984) had been published.

The strategy finally adopted by SIGMA was therefore to structure respondent selection round the two factors known to maximize variation in homosexual behavior:

- Age, and
- Type of Relationship.

Enter the network, via Rapoport (1953, 1957), Rapoport and Horvath (1961), Fararo and Sunshine (1964)—and snowballing.

1.1. Sampling Via a Large Social Network: Theoretical Ideas

The conceptual basis of the sampling strategy was motivated by the question of how to obtain systematic (ultimately unbiased) information about a large, unknown, and connected social network of sexually active gay and bisexual men. I was familiar through earlier sociometric interests with Rapoport's models of diffusion through large networks. His original application in mathematical biophysics had been to the neural net, estimating the "gross statistical properties" (Rapoport 1963: 512) of a huge network of

unknown size and structure, and in particular the "close-knittedness" and the "ultimate connectivity" of the neural net.[3] This approach was applied in turn to communication and (by Fararo) to friendships among young and institutionalized offenders.

The aim in these social studies was to explain diffusion within these empirical social networks, the manifest divergence of these processes from diffusion in a purely *random* baseline network, and to do so by identifying and estimating significant *bias parameters* (Rapoport 1957, Fararo and Sunshine 1964, Skvoretz 1985). These parameters, when incorporated into the "reduced axone density" (in effect: the "slowed-down" effective outdegree, referred to as a), explained the (lower) ultimate connectivity and rate of increase in new contacts by means of such modifications of a random Polya model.

In this model the expected fraction contacted at time t is given by the iterative equation (Rapoport 1961:285):

$$p_{t+1} = (1 - X_t)\ (1 - e^{-ap_t})$$

where p_t denotes the expected proportion of new contacts at step t in a random net, and X_t the cumulative proportion contacted at time t. The expected ultimate connectivity X_∞ satisfies the transcendental equation:

$$X_\infty = 1 - (1 - p_0)\, e^{-aX_\infty}$$

where X_∞ is the asymptotic value of the cumulative fraction contacted. The epidemiological parallel is of course obvious, and in later work Rapoport actually referred to the model as the (Polya) "contagion process model" (Rapoport 1979).

The method for making the estimates of diffusion or "infection" is the "tracing," which consists of producing a rooted tree giving the new contacts at each ordinal step t. Starting at a small initial fraction of nodes, the tree is "grown" by consulting the sociometric contact matrix at each new contact, until no new contacts occur. A small illustratory example is useful.

1.1.0. The Tracing

Let us suppose that the network structure is known (which is often not the case, but simplifies the exposition), and is given by the adjacency matrix **A**, whose element a_{ij} is 1 if there is a relational link between points i and j and is 0 otherwise. An example is given in Figure One(A). Here the number of points, N, is 10 and the outdegree of each point (number of contacts made) is a constant 2. The *starting set* for each tracing is always small compared to N, and in the two examples shown it consists of a single point: point E in Figure One(C)(i) and point E in Figure One(C)(ii).

	A	B	C	D	E	F	G	H	I	J
A	0	0	1	1	0	0	0	0	0	0
B	0	0	1	1	0	0	0	0	0	0
C	0	1	0	0	0	1	0	0	0	0
D	0	0	1	0	1	0	0	0	0	0
E	0	0	0	0	0	1	0	0	0	1
F	0	1	0	0	0	0	0	1	0	0
G	0	0	1	0	0	0	0	0	1	0
H	0	0	1	0	0	1	0	0	0	0
I	0	0	1	0	0	0	1	0	0	0
J	1	0	1	0	0	0	0	0	0	0

Figure One (A)
Network Adjacency Matrix, N = 10, a = 2

Step	**0**	**1**	**2**	**3**	**4**	**5**	**6**
p (t)	.1	.2	.24	.16	.06	.05	.03
X (t)	.1	.3	.54	.70	.76	.81	.84

p (t) proportion of new points contactd at step *t*
X (t) Cumulative proprtion of new points contactd by step *t*

Figure One (B)
Network Tracing Distributions Averaged over 10 Starting Sets
(each point taken in turn)

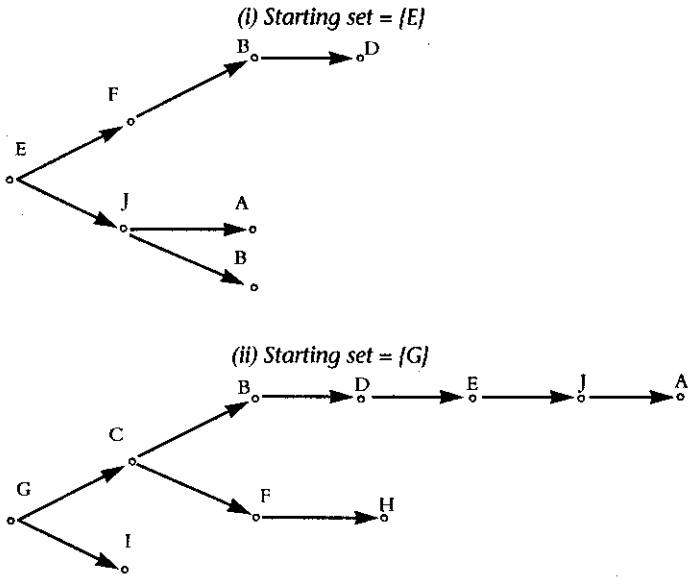

Figure One (C)
Illustratory Network Tracings

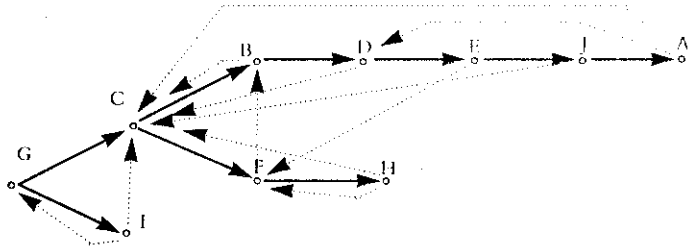

Figure One (C)
Illustratory Network Tracings (cont.)

Each tracing proceeds by defining the starting set (step $t = 0$) and then noting the contacts which this starting set makes. These constitute the set of (two) new points contacted at step $t = 1$, and these are counted into n(1) (the number of new contacts at step t.= 1). The contacts of each of these new contacts in turn are now identified, but only if they have not been contacted previously do they count in n(2). This tracing process continues until there are no new contacts; this stage of "ultimate connectivity" may be complete (contacting the entire set of 10 points as in tracing (ii)) or partial (as in tracing (iii) in Figure One (C)). Notice that the number of points newly contacted at each step and the number of steps taken to reach the end of the tracing differs depending on the starting set.

To emphasize the fact that it is *new* contacts only that are considered at each step, tracing (ii) starting with point E is represented in tracing (iii) by drawing in the redundant (already contacted) links at each step as a dotted arrow. Because they have already been contacted, they of course point backwards.

A given tracing is characterized by the fraction of the population finally contacted (average ultimate connectivity) and by the number of steps taken to reach this stage. In Rapoport's theory, a number of tracings are made, until the *average* tracing distribution (i.e., averaged at each step over all the tracings) stabilizes to within a desired limit. This asymptotic tracing distribution is given in Figure One(B). The *proportion* of new points contacted at each step is (simply $n(t)/N$) is given in $p(t)$, and the *cumulative* proportion of points contacted by step t is given in $X(t)$. The "ultimate connectivity" of the network is given by the last entry. Thus *on average*, eighty-four percent of the points are ultimately contacted by the sixth step when a tracing is made in this network.

The interesting difference between the neurological and the normal social science applications is that tracings are usually produced in the social

science case by referring to a *known* sociomatrix (i.e., relating to a population of known size and contacts), and hence what is constructed by the technique is a *synthetic* tracing. By contrast, the original mechanism proposed by Rapoport referred to a network of unknown size, and used the empirical physical process of exciting the initial neurone/s and observing the pattern of firing across neurones. These then produced estimates of the size and connectivity of the population. Paradoxically, the use of Rapoport's techniques in the case of tracing hidden populations thus means reverting to his original conceptualization.

1.1.1. Sociological Implementations

The conceptual analogy with sampling gay men seemed well-nigh complete. Implementing a tracing (or several) provides an excellent methodological specification of what sampling a hidden population should be. If continued to completion (and without error) such tracings would provide an enumeration (of at least connected subsets) of the homosexual population and also information about its local network characteristics. This is, of course, an ideal type and is practically unrealistic as a technique *in toto*. Nonetheless, it tells us what "snowball sampling" should be. A close parallel to the tracing process was known to sociologists from Coleman's study of diffusion (Coleman 1958), where physicians were asked to name those of their colleagues to whom they gave information about a new drug, a method now known as the "chain-referral" method (e.g. Biernacki and Waldorf 1981).

But there are important differences. In the sociological examples, attention has shifted from Rapoport's interest in the inherent properties of the network and the mechanism generating it to the usefulness of a *strategy* for reaching hidden or hard-to-reach populations. To be sure, the underlying assumption is that there exists a network of contacts (for otherwise why would snowballing work?), but even considerations like the number of stages/steps and the outdegree (constant or average number of contacts) slips out of sight.

Paradoxically, most statistical work on snowball sampling (Goodman 1961, Holland and Leinhardt 1979, van Metter 1990, Snijders 1992) has not generally been directly relevant since it typically refers not to populations of unknown size, but rather to issues of inferring population network properties from samples. Few indeed are the studies claiming snowball status which come in any way near satisfying requirements such as sampling to exhaustion. Perhaps, given the wider meaning nowadays given to snowball sampling—in effect, the cumulative but haphazard acquisition of a quota or convenience sample—it is better to use the term "tracing sample" to refer to tracing procedures which follow contacts-of-contacts to the exhaustion of new contacts.

1.1.2. Sampling Gay Men: SIGMA's Sampling Procedure

The sampling process used in SIGMA was two-stage: first to obtain easily-accessible respondents in each of the nine Project Design typology cells (chiefly from gay pubs, clubs and voluntary organizations);[4] secondly to use these initial contacts as starting samples for producing tracing trees. The imagery sometimes used in SIGMA to refer to this process was "burrowing into the iceberg"; indeed, given the fact that those most "out" are a highly biased group representing only the tip of the iceberg, the idea of obtaining less "out" contacts of the same Age x Relationship-type by snowballing is attractive. But notice the looseness of the relational definition we used; in practice the interviewer asked the initial respondents to name other potential respondents who were of the same (Project) type as themselves, but preferably less "out" as gay. It was left to the interviewer to satisfy him/herself that this definition was understood by the respondent, and we were rarely able to ascertain whether this had actually been done.

The attempt by SIGMA to implement tracing sampling was noble, but ultimately deficient, and for a number of instructive reasons:

- often a given gay man's friends and acquaintances are *not* of the same Age-Relationship type as himself, so that it was frequently quite difficult for a respondent to name someone of the same type, let alone someone that was less "out."
- the *number* of contacts to be named was never specified; more relevantly, there was no criterion provided by which the respondent could decide when the number of his nominees was sufficient.
- as a Project we had bound ourselves to anonymity in the form of not recording or making use of the name of anyone named in the research context. We therefore had to rely upon the respondent to contact his nominee and ask him to participate in the Project. Consequently we might never know that a specific person had been thus nominated, let alone whose nominee he was.[5]

However, in terms of the stated objectives—to "snowball" into the more covert gay population—there was some degree of success. The first "Question Schedule" contained a number of questions asking who knew that the respondent was gay/bisexual. Inter alia this provided a useful indicator of "outness," and (at least in the South Wales site) this index of "outness" decreased as known contacts were interviewed.

But it must be said that the exercise was not overall a resounding success, and that to all intents the initial SIGMA sample was no more (nor indeed any less) a "snowball" than other such studies. The reasons cited above are enough to account for its lack of success, but in principle each could be

remedied. In Section Three, I present information about how critical the actual naming (identification) process is. But the main shortcoming was that the criterion/relation for respondent naming was not only too vague, it was also not related directly enough to the sexual transmission method we were studying. This raises the question of whether a genuine (sexual) tracing sampling technique could have been devised and implemented, and whether it would have been more relevant. Without doubt the attempt might bring to light yet more compelling shortcomings (not least those connected with confidentiality and compliance) and would probably be more expensive in terms of time and money. I shall return to this in the next sections; it remains to assess the depth of the burrowing.

The gay scene in Cardiff and area is a good deal more closely-knit (on any significant criterion) than that of the sister site of London.[6] In Rapoport/Fararo terms, this implies higher values of reciprocity ("sibling bias, s") and transitivity ("parent bias", p).[7] This in turn implies longer chains because the clustering leads to more redundant (already contacted) new contacts at any step, and thus "slow down" the growth of X(t) and of the eventual connectivity asymptote.

Where it was possible to track the contacting process in Cardiff, these redundant new nominations occurred with sufficient frequency that a rule was drawn up that new nominations had to be checked by the site office before being allowed as a new sample member; in London this rarely happened.

For sexual contacts (of whatever variety) there turned out to be a goodly number of cross-cutting circles, but with weak links between them, so that an estimate of ultimate connectivity probably depends rather importantly on whether the sample includes the liaison persons (bridges) that mediate such clusters. Not including bridges would lead to a falsely low estimate of ultimate connectivity (and hence of prevalence). It would also lead to missing certain important subsets of respondents who come in and out of the scene on an occasional basis and who would only normally be contacted via one man; occasional (but not hardened) users of "cottages" are an important example, and a relevant important sub-group epidemiologically.[8]

These assessments of the topology of the homosexual network are largely impressionistic, and would need to be investigated directly as hypotheses. In neither site, however, did we normally exceed a chain-length (let alone a tracing step-length) of more than three, and we argued that a length of four would be necessary to achieve even reasonable coverage, and ten or more would be necessary to come anywhere near encompassing a coherent cluster (Davies 1986).

The question of whether engaging in tracing samples of actual sexual

contact (and especially of implicated behaviors such as anal intercourse) would materially improve our network-knowledge and assist in estimating prevalence and prediction is a good deal more moot, and brings us nicely to the second section.

2. Networks of Transmission

The main implicated behavior for the sexual transmission of HIV among gay men is, of course, anal intercourse, with the remoter (and contested) possibility of fellatio (but see Koopman et al. 1992). An important framework for interpreting and predicting transmission is therefore the *sexually defined* network. The defining relation can be sexually multiplex (most gay men don't just have anal intercourse) and will normally need to be restricted to a fixed incidence time-period. The sexual relation is also necessarily asymmetric, since the probability of transmission is greater for the passive partner (Darrow et al. 1983, inter alia).[9] This asymmetry means that sexual networks and tracing trees have to rely on directed graphs. Moreover, sexual role, though normally considered as an individual or point property (see Coxon et al. 1993) can more appropriately be viewed as a dyadic property. Thus, although in individual terms a man may engage in only *A*ctive, only *P*assive, *B*oth (or *N*either) modality (the so-called *BAPN* roles), this is often distributed associatively so that a man is (say) only active with one partner and only passive with another (this will still lead to his being classified as "both" in the individual-based role categorization).

It is important to specify why the Rapoport model is being used as a basis. After all, HIV is not generally highly infective, and one contact is rarely sufficient for infection.[10] We also know that epidemiologically it is necessary to take into account the stage of HIV infection in assigning probabilities of infection. What the Rapoport models do is to provide us with a *potential* infection model—a network structure of routes and paths along which infection can travel, and where (as with a block-model) the zeroes are often as significant as the ones.[11]

Many existing epidemiological models are long on necessary conditions for infection but short indeed on the socio-sexual structure which in large part constrains and contains the infection. Equally, these models often make demands for data that normally are not collected, particularly by survey analysts—such as information on mixing (Anderson et al. 1986) and especially on the second-step information on the number of sexual partners of one's sexual partners. These become crucial parameters in the Anderson model.

2.1. Piloting a Study of a Two-Zone Network of Anal Intercourse

At an early stage in the pilot work of SIGMA in 1984 (reported in Coxon

1986) I began investigating the possibility and viability of mounting a network study of anal intercourse within the South Wales SIGMA site. This would be akin to contact-tracing of STD Clinics but more sensitive and sociologically useful. It is relatively straightforward to obtain information about whether or not a gay man engages in anal intercourse, the rates and number of partners in a given incidence period, and the extent to which condoms are used. These are now routine questions in any socio-epidemiological enquiry. Only rarely, however, are the pieces put together so that, for instance, we know whether anal intercourse is engaged in with a particular partner, and with what modalities.

In Project SIGMA we have developed the Sexual Diary as a supplementary method for obtaining precisely this information (Coxon 1988; Coxon et al. 1992, 1993; Davies 1990). The main shortcoming is that by the self-denying ordinance mentioned above, we specifically have *not* asked for the names of partners (though we do ask for their descriptive attributes), thereby apparently foregoing the possibility of linking the data to obtain sexual networks.[12]

The other strategy was to attempt to construct an anal intercourse network directly, by following a proper tracing procedure, initially with a single root sample. In this example I refer to one such network, but restrict attention to the initial one-step partners of Respondent #One, although the tracing actually continued beyond this point. The tracing procedure was as follows:

Sex Tracing

0 specify the incidence period (and bounds)
1 choose the root node/s
2 establish how many partners he had engaged in anal intercourse, get their name, information about them, details of act/s of anal intercourse, if possible, including modality, ejaculation and condom use
3 for each named partner, repeat step 2.

(The sociogram of the network centered on Respondent One is given in Figure Two.)

Step 0 This is straightforward. In the example it was taken as a six month period, and limited to the South Wales Project area.

Step 1 This starting node is chosen for good (cautionary) reasons. His position is discussed below.

Step 2 It is fairly unusual for a gay man to have thirteen partners with whom he engages in anal intercourse, but by no means rare. It was not difficult to establish information about the initial thirteen linked partners from him, and his account could be checked against his Sexual Diary account. At the level of whether or not he had had anal intercourse, and in what modality, the two accounts concurred.

Step 3 Knowing the identity of the partners, it was possible to match ten of the partners with SIGMA respondents (all except Respondents Twelve, Fourteen, and Four), and it would have been possible to find the inter-zone One links from Diaries or Interviews and not have to rely on contacting them directly. Information could *only* be obtained for the three non-SIGMA respondents by direct contact, and in one case at least (Twelve) we knew that he had already refused to be a part of the Project panel, and hence was highly unlikely to give information on his anal intercourse partners. Beyond zone-One (i.e. three-step and beyond) I have only indicated contacts *outside* the initial thirteen.

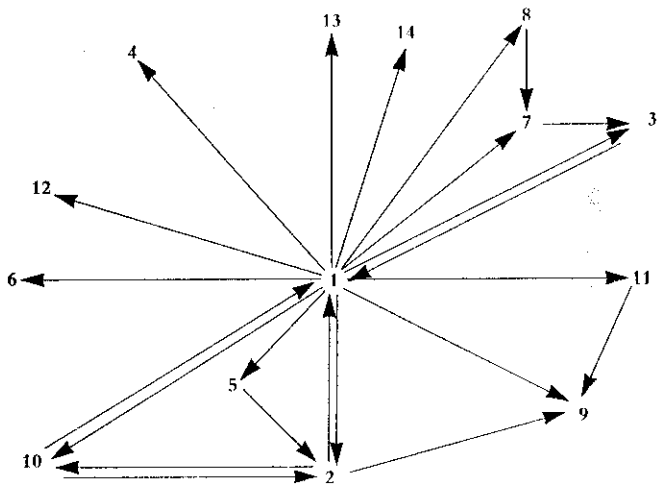

Figure Two: [1] Ego-centric Graph of Anal Intercourse

Several comments are in order:

- The root node (1) is a professional man in his mid-thirties, with two regular partners (two and three); six, eight and ten are more-or-less regular partners—"occasionals" might be a better term. Partner Five is originally an "affair" (regular partner) of Two's. The rest are mostly casual partners, but include a few more-than-once casuals.
- In terms of individual sex-role, almost all of the men are "B" [both active and passive] during this period, but Six is A[ctive only] and Nine is P[assive only].[13]
- In terms of the sociometric structure,[14] the core "clique" (maximally

connected symmetric subgraph) consists of {1,2,10}, with {3} as an important appendage. Partner Nine is a particularly interesting (and at-risk) person, receiving by far the highest number of direct and indirect paths of anal intercourse, whilst Ten by contrast *generates* the maximum number of anal intercourse paths (see Appendix 1).

- In terms of the Rapoport bias parameters, reciprocity is 0.17 (fairly close to the proportion of "Both" in larger studies [Coxon et al. 1993]), and transitivity ("parent bias") is 0.04, which is naturally attenuated in this restricted set, but which is not much less than the value for larger networks—after all, anal intercourse is only likely to be (trivially) transitive among those who are both active and passive (e.g. {1,2,10}), though "genuine" (transitive in one direction) chains do appear, as in {1,7,8}).
- In 1985 (to which the data refer) there is *no instance at all* of a condom being used in this group during the six-month period, but the HIV prevalence rate was also very low in this non-Clinic South Wales subsample (Hunt et al. 1990).
- This small world is by no means self-contained. At the second, and especially at the third step, the contacts begin to leave South Wales, and also include female partners. Partners {10, 3, 13} have a female partner, and {5, 3, 7, 6} have one or more male partners in London, which has always had a higher HIV sero-prevalence rate than South Wales.

It is very dangerous to over-interpret these small and illustrative data; the point of the exercise has been to show that it is a potentially valuable exercise, albeit a difficult one. In particular, it involves considerable data-collection costs and requires extraordinarily sensitive and delicate skills on the part of the interviewer.

With the adoption of the self-denying ordinance in not recording or using names, the attempt to implement anal intercourse tracing was virtually abandoned, not least because the funding agency disallowed it—on financial, not ethical grounds, it should be said.[15] So: was this another resounding non-success? Let us see.

2.2. Inferring Partner Identity

Perhaps the most galling thing about the ordinance about anonymity was that in many instances, respondents *did* give names of partners. The reluctance was not because they did not want to name partners (at least under conditions of confidentiality), so much as their concern that the nominee would find out that he had been named by the respondent. But even when the respondent did not name his partner/s it was sometimes possible to infer his (or her) identity from circumstantial or public knowledge—interviewers themselves were

usually involved in the local gay scene—and even from within the interview. The accuracy of such inferences and guesses was much aided by the fact that although *names* were not asked, descriptive information was.

2.2.1. Partner Characteristics: Matching Attributes

From the Wave One "Question Schedule" (1987) onwards, respondents were asked to give a range of attributes for their regular partner/s—Sex, Age, Race, Job, Marital Status and Domicile—and throughout the Project all Sexual Diary keepers are asked to describe each partner during the month in terms of the following characteristics:[16]

Partner Attributes (Sexual Diaries)

0 Partner (sequential number and) Sex
1 Status (Regular, Occasional, Casual/One-off)
2 Age (known or guessed)
3 How long you've been having sex
4 Where met (on this occasion) [Casual partners only]
5 Other information—Job, basis of attraction, payment?
6 HIV status, if known

Members of Professor Roy Anderson's team at [then] Imperial College had expressed interest in using the Sexual Diary data for making estimates of mixing ratios—i.e., of information not only about the number of partners, but also of the number of partners of partners.[17] If partners were identified by name (and, more demandingly, if they were Project diary-keepers themselves), then such a procedure of estimation depends simply on establishing linkage between a respondent's file and those of his partners. But for reasons explained in the last section, the SIGMA Undertaking on Confidentiality meant that this was not possible. Dr. Chris Joyce originally suggested a possible strategy: to use information which a diarist gave about a given partner as a yardstick or template "profile" and then attempt to identify him by searching for (preferably one) respondent whose data matched that information.

One option (not systematically followed out) was simply to use the partner attributes, so that if, for example, Jim had described his Partner Five as:

> *A regular partner / aged 24 / with whom I've been having sex for 2 years / who is a well-endowed teacher from Manchester / and HIV antibody negative*

then a SIGMA diarist, Fred, who was found to have the same characteristics would be identified as putative Partner Five, and the information he gave about the number of his sexual partners would then be derivable from his Diary record. The main difficulties are:

- What counts as "the same characteristics," especially given that the fourth category ("Other information") is open-ended?
- Will a pair describe themselves in the same way? (Thus, Jim's "regular" partner Fred might consider himself an "occasional" partner of Jim.)
- How much "noise" / tolerance for error can be allowed in matching, especially given the fact that reported age is often systematically biased downwards?
- What happens if more than one candidate appears? Suppose Fred1 is *"Regular / 28 / sex for 2 years / student from Stockport / HIV neg."* and Fred2 is *"Occasional / 26 / sex for 2 years / student at Salford University / HIV neg"*; which shall be identified as the "real" Fred?[18]
- In the case of casual or one-off partners, it is quite likely that information will be very deficient and grossly misperceived, making matching virtually impossible. And yet these cases may often be very important epidemiologically.

Matching by attributes (and especially on open-ended and multi-reference ones) can therefore be a very hazardous procedure, and if there are too many mismatches then the subsequent constructed network will become almost misleading and very possibly useless.

But all was not lost; rather than tighten up or extend the partner characteristics, a different tack was employed.

2.2.2 Retrospective Networking

The next ploy was to ignore such individual characteristics (at least initially) and concentrate on much more detailed, relevant and less-easily matchable data. This consisted of "retrospective networking"—namely matching a partner by matching the *sexual session* itself in which they had both participated (after the event, an obvious choice to make!).[19]

The process had three steps: *screening, temporal matching, and behavioral matching.*

2.2.2.1 Screening

First, all solitary sessions (involving only one partner, i.e. the diarist) were removed, which amount to exactly one-third of the sessions: 33.3 percent (Coxon 1990:15) in this set of data.[20]

2.2.2.2 Temporal matching

Obviously, a match should only occur if the sessions involved had taken place at the same date and time. However, some leeway had to be allowed on this, since earlier studies had shown that displacements up to one day between the

partners' accounts were not unusual. The matches produced by this process were then treated as fulfilling a necessary condition for matching.

The results of purely *temporal* matches are presented in Figure Three.

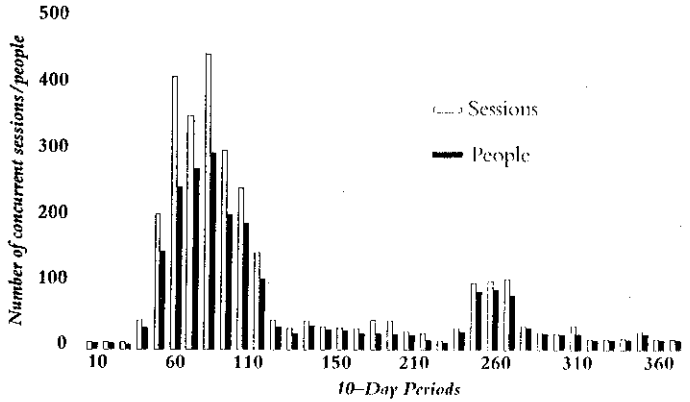

Figure Three: Number of Temporally Matched Sessions and People Over Diaries Epoch

- First, all sexual sessions involving only one person (e.g., solitary masturbation) were excluded, leaving 2783 sessions for 166 diarists.
- These sexual sessions occurred on a total of 359 days, with an average of 7.6 per day (and a maximum of 66).
- The maximum number of diarists reporting one or more session in any one day is forty, with an average of 5.4 diarists/day.
- A possible match consists of any *pair* of sessions which occurred on the same day and which could therefore represent the same session written from each partner's viewpoint. . .
- . . . but this means a large number of potential matches: for these data, either 15,938 matched sessions, or 14,374 potential partners; a truly impressive, but unrealistically large, number.

2.2.2.3. Behavioral Matching

Having established a potential match by temporal matching, the next stage consisted of matching a given session and a potential identically referring (matching) session by comparing the strings (encoded sequence of sex acts) embodying that session.

Before explaining this process, a brief excursus on the encoding of sexual behavior is in order (see Coxon et al. 1992).

2.2.2.3.1 Encoding the Sexual Session as a String

Diaries are initially written in everyday language according to a specified

structure (see above). On receipt they are then encoded into a more efficient form as a quasi-linguistic string (Coxon et al. 1992). An example follows:

Diarist's Account:

We'd been drinking in a gay bar in Amsterdam, where we met. He was 30s, from Utrecht, into leather. We went back to my hotel room just after 1 a.m. and after sharing a joint, went to bed. . .

First we kissed, then I sucked him and he then sucked me and after that we moved into a '69' position and he came in my mouth. Then I moved to fuck him and he put a condom on me with KY over it. I entered him (him sitting on me) and he was wanking himself, and we started using poppers and I then came in him and then he came over my chest. After pulling out, we kissed for a while then went to sleep.

Encoding of Account
(PARTNER LIST) C3, 30s, male, Utrecht, into leather
(ANTECEDENTS) Met in gay bar, Amsterdam, drinking; shared marijuana. My hotel room, 13.00 26/8/88.

(encoded string of sex act sequence):
MDK AS PS MS,NM (AF,CN/l&HW,NI)/p MDK

The extended natural language description of the session is thus reduced to the code string (last line), used in all analysis programs.

There are several ways of now performing such a matching; one consists of defining a Levenshtein distance between two sequences (Sankoff and Kruskal 1983:18 et ff), but in this instance a related method was used which was developed in molecular biology for sequence homology searches (Feng et al. 1985).[21] These methods produce a score by which a match could be said to be made if it exceeds a stipulated limit. In brief, the sequence of sexual acts making up a session of the reference subject should be the same as the potential partner's session if it is to match. There are some difficulties and provisos which must be made. First, the respective *modalities* for any act must correspond rather than be identical (e.g., if the reference subject is *active* for a given act, the partner must be *passive* for it). Moreover, allowance had to be made for "chunking": some subjects make finer distinctions than others in describing a sequence. Finally, a "rarity" weighting was applied; some acts (e.g. "fisting," ano-brachial insertion) are very infrequent, and their occurrence therefore has a higher surprisal value.[22] Their occurrence is deemed more important in matching than common acts such as masturbation.

When comparison was also made of the content of the strings, the rate of matching reduced considerably: now only two percent of strings are matched as referring to the "same" session.[23]

Further descriptive information in the diary entries (e.g., the occupation

of the partner, or whether drugs were used) can tighten up yet further the likelihood of partner identification, though this has not yet been done to any considerable extent.

2.2.2.4. Evaluating "Anonymous" Identification

At this juncture, little more can be said about the validity of the process: the matching certainly has face validity, but we cannot know that (for this data set) the match is correct, since it is a totally anonymous (unidentified) set of data. The next stage of validation will consist of testing whether the procedure can identify the (known) real partners from their session strings. In the meantime an interesting and potentially very important hypothesis arises concerning the "class of potential partners" in a known sexual session. Quite independently of the question of how to find the "real" partner—but assuming that s/he is in this class—how similar or homogeneous is the subset of candidates for being the partner? Do those engaging in the same (or structurally identical, or similar) sequences of sexual behavior resemble each other in other ways? In particular, is their pattern of other partners similar? If so, this will provide confirmation of the "like-me" characteristics of other social networks, and incidentally provide aggregate estimates of mixing for the Anderson models.[24]

In the meantime there are two ways in which this technique can lead to improved and more extensive matching:

- *"Coverage" should be as high as possible:* even when a match occurs there is a possibility that it is fallacious in the sense that partners *outside* the diary sample were actually involved and the temporal and behavioral matching was therefore purely coincidental. With higher proportions of the sexually active actually involved, this probability is decreased.
- *The date-limits of the diaries, should be as close to identical as possible,* i.e., if most respondents are completing their diaries over the same period, there is again a higher probability of correct matching.

Current research on the reliability and validity of the diary method and its use in this context is also helping to us to understand how and when two scripts or session-strings are to be considered identical.

3. Conclusions

In Project SIGMA we have tried at all times to keep the atom of social networks—the dyad—as our main unit, rather than the individual. Most obviously this means looking at (non-solitary) sexual behavior as being primarily the outcome of an interaction rather than as an individual propensity, as processes such as negotiation then become central to explaining what

is occurring epidemiologically. The recent debate on so-called "relapse" provides a good instance of the difference that this makes in interpretation.

But the larger issues of networks also loom large, and we have to admit mixed success. This chapter has therefore been written in a self-critical mode, so that others can learn from our mistakes and successes. Given the unwillingness of funding agencies to underwrite some of the more ambitious proposals (very possibly with good reason), we have had to take a more pragmatic line than perhaps we would have wished. But the study has also shown us that a surprising amount can be done within the confines of a fairly conventional longitudinal study to examine network characteristics. For example, examination of triads of sexual and other interactions allows direct estimates to be made of transitivity and other "bias" parameters even if we know little about the detailed topology of the entire network. But equally there are some matters which, however difficult and expensive, can only be tackled by a direct, full-blown network methodology. It will be a tragedy if such work has to be abandoned in favor of a purely individual-based mode of enquiry and analysis which cannot represent these crucial aspects of transmission.

Notes

1. I am grateful to the Department of Health and to the Medical Research Council funding which supports the work reported here; the views expressed are my own and do not necessarily represent those of the funding authorities. I am also especially grateful to Dr. Christopher Joyce (now Research Scientist, AIDS/HIV Division, Communicable Diseases Research Centre, Colindale, London) for his collaboration and work represented in section 2.2 of this paper.

2. The acronym SIGMA represents Socio-sexual Investigations of Gay Men and AIDS. Project SIGMA is a is a longitudinal, non-clinic based, serological and behavioral study of the sexual and social lifestyle of gay and bisexual men in England and Wales. (It is also part of the English study under the auspices of WHO Global Programme on AIDS Homosexual Response Studies).

   SIGMA is one of the largest cohort studies in Europe and the only study in the U.K. to have emerged from the gay community. Initial work began in 1983, and funding followed in 1987. To date, the Project has interviewed over one thousand men, half of whom have been interviewed four times at [median] intervals of ten months. The main aims of the study are to describe the sexual behavior and lifestyles of gay and bisexual men; to monitor changes in sexual behavior in relation to HIV/AIDS; to examine attitudes to different sexual behaviors and relationships; to investigate reactions to safer sex practices; to estimate prevalence of HIV and other viral infections in a non-clinic group of gay and bisexual men.

   Project SIGMA uses several complementary methods of obtaining information, including:

   - The detailed structured interview in which each respondent is asked for detailed information on sexual history and current practices (centered upon the Index of Sexual Behavior [Coxon et al. 1992]), numbers and characteristics (but not names) of sexual partners, health, and attitudes towards HIV and safer sex.

   - The sexual diaries (Coxon 1988) are a daily record of sexual activity kept by respondents for a month after each interview. So far we have collected information on about thirty thousand sexual encounters which allows a unique analysis of their structure.
   - Blood and/or saliva samples are also collected at the interview by trained staff and tested for HIV-1 antibodies and other viral markers. Results are available to respondents through trained counsellors.
   - The postal survey of sexual behaviour is a self-completion questionnaire which appears in the gay press periodically.

3. "Close-knittedness" (although not Rapoport's term) is measured by the average rate of acquisition of new contacts, and the number of steps (links) necessary to reach exhaustion—the latter referred to as the "ultimate connectivity," the fraction of the population finally reached.

4. The two factors of Age and Relationship-type were trichotomized and crossed to produce the nine-fold typology. Age was split at twenty-one (the age of homosexual consent in England and Wales) and thirty-nine (after which men would have grown up when any homosexual behavior was illegal, before the 1967 Sexual Offences Act), and Relationship-type into Closed, Open and No Regular Relationship.

5. At a later stage we decided that the anonymity undertaking might be a case of shooting ourselves in the methodological foot (see the discussion in Coxon et al. 1993), but there were (and are) excellent reasons why gay men need to be persuaded that such information is safe and cannot be used against them.

6. Defined as the old county boundaries of the County of Glamorgan, but perhaps more accurately expressed as "Caerdydd a'r cylch"—Cardiff and the surrounding area.

7. **"Parent" (p) bias:** $\mathbf{Pr\ \{xRy \mid yRx\}}$**, and**

   "Sibling" (s) bias: $\mathbf{Pr\ \{xRy \mid \exists\ z,\ zRx \wedge zRy\}}$.

   These combine with the actual outdegree ("axone density", a) to form the "Reduced axone density", a) by the following expression:

$$a = a - p - (a - 1)\,s$$

8. A.k.a. "tearooms" (USA) and "beats" (Australia), i.e., public toilets. Sub-projects of SIGMA and of the St. Mary's Paddington group studied this group in various locations—though many aspects were at variance with Humphreys' (1970) conclusions, especially about the considerable preponderance of heterosexually married men frequenting the cottages (though they are a significant fraction.) Although the sex which takes place is by and large safe, there are particular (often more deserted) ones where entirely unsafe sex is the norm.

9. Terminology referring to the modality of sexual behavior is confusing (see Coxon 1988; Coxon et al. 1992). Common medical usage is "inserter" vs. "insertee" (but this is useless for non-penetrative sexual behavior, such as masturbation); and "donor" vs. "recipient" assumes ejaculation. We prefer "active" vs. "passive," which is unambiguous and in common parlance among gay men. Following linguistic usage, the active partner *does X* to the other partner whilst the passive partner *has X done to him* by the other partner. Note that "inserter" is not necessarily equivalent to the "active" partner.

10. Present research suggests that it may well be highly infectious at particular stages, especially after infection and before sero-conversion; more worryingly, the same may be true of fellatio at this stage (see Koopman et al. 1992).

11. I.e. the "never fuck" category (0 link) is often more stable over time than the other individual roles (Coxon et al. 1993).

12. This shortcoming and ways to attempt to overcome it are described in Section Three.

13. Respondent fifteen is P as regards anal intercourse with men, but has active vaginal intercourse with a partner in London.

14. Analysis using UCINET-IV (v1.02) programs (Borgatti, Everett and Freeman 1992).

15. At later waves, this prohibition was relaxed, though not to any considerable extent, and it is still under discussion in the Project.

16. Together with information about the partnership—how long known, sexual in/exclusivity, number and sex of any other partners of the partner, and frequency of seeing.

17. The sexual diary data used were those obtained from the first *Gay Times* cohort, collected in 1986 and documented in Coxon 1990. The total data-set refers to 188 individuals, who performed 7532 sexual acts in an average of 23.4 sessions in the month.

18. The problem is not unusual in any matching context, but five characteristics of this sort are not usually enough fully to disambiguate different candidates.

19. A "sexual session" in the structure of sexual behavior developed by Project SIGMA (Coxon et al. 1992) is the main unit of such behavior: the "sentence" of sexual activity—self-sufficient and intrinsically well-formed. It occurs at a given (specified) time and place, consists of one or more sexual acts, and involves at least one, usually two (and sometimes more) people. It is typically terminated by sleep, a non-sexual intermission, or change of partner. The characteristics of a Sexual Session include four components: the *Setting*, the *Precedents* [e.g., prefatory drink, nitrites], the *Accompaniments* ["toys," leather], and the *Partner Specification* mentioned above, and the main content is the sequential specification of the components' sexual acts.

20. This figure plus or minus five percent holds for all the sexual diary data sets.

21. Which basically estimates the number of insertions and deletions necessary to turn one string into another.

22. A term borrowed from information theory (Luce 1960), where the "surprisal" of an act is inversely proportional to the frequency of its occurrence: $h_i = log_2 p_i$.

23. A total of 344 possible matches involving 298 individuals—i.e., a possible matching of 149 pairs of individuals from a total of 166 diarists.

24. Similar concerns, with a different methodology, have been expressed in the context of research on AIDS in Africa by Orubuloye, Caldwell, and Caldwell 1992.

Appendix One
Volume of Direct and Indirect Paths Between Points in Figure 1

PTS:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	3	2	1	1	1	1	2	5	3	1	1	1	1
2	2	0	4	2	2	2	2	4	5	2	2	2	2	2
3	1	3	0	1	1	1	1	2	5	3	1	1	1	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	2	1	4	2	0	2	2	4	5	2	2	2	2	2
6	1	3	2	1	1	0	1	2	5	3	1	1	1	1
7	1	3	1	1	1	1	0	2	5	3	1	1	1	1
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	2	3	4	2	2	2	2	4	7	0	2	2	2	2
11	0	0	0	0	0	0	0	0	1	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0