HOMELESS OUTREACH IN THE TRI CITIES: IS SOMETHING SOCIAL GOING ON?

by

Laurens Bakker
B.Sc., Roosevelt Academy, 2008

A Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
in the School
of
Computing Science

SIMON FRASER UNIVERSITY
Spring 2011

APPROVAL

Name: Laurens Bakker

Degree: Master of Science

Title of Thesis: Homeless outreach in the Tri Cities: is something social going on?

Examining Committee: Dr. F. Warren Burton
Chair

Dr. Binay K. Bhattacharya, Professor
Computing Science, Simon Fraser University
Senior Supervisor

Dr. Patricia Brantingham, Professor
Criminology, Simon Fraser University
Senior Supervisor

Dr. Vahid Dabbaghian, Adjunct Professor
Mathematics, Simon Fraser University
Supervisor

Dr. Nilesh Saraf, Assistant Professor
Business, Simon Fraser University
Examiner

Date Approved: 28 April 2011

Declaration of
Partial Copyright Licence

Abstract

Societal perception of homelessness has shifted from it being an individual to a social problem, reflected in the changing interpretation of "home:" from 'house' to 'place of belonging.' Although appropriate, the broadened definition poses a challenge in measuring the outcome of homeless outreach: provision of housing can be counted whereas provision of belonging cannot.

Data collected by the Hope for Freedom Society, a homeless outreach organisation in the Tri Cities, were explored in search for evidence of social interaction and belonging. We anonymised the provided raw text activity logs and extracted a network containing clients, locations and dates. Projections of this network were then used to examine community structure in the clients' social network and activity space, using statistical models and modularity-based community finding algorithms.

Despite the inclusion of interaction information, personal information, locations and dates, we were unable to find evidence of social interaction in the data.

"If men define situations as real, they are real in their consequences."

William Isaac Thomas

Acknowledgments

First and foremost thanks goes to the MoCSSy members,[1] past and present, for a great productive time taking advantage of people, dispensing herring and high-calorie, high-caffeine bribes.[2] I would specifically like to thank Warren Hare, who showed me what good modelling is about, and Philippe Giabbanelli for countless pointers to relevant references, blunt but constructive criticism, and a motivating example. Just as valuable were the people who asked myriad challenging questions: Azadeh Alimadad, Afsaneh Bakhtiari and Mona Vajihollahi. Without all of these people, I could not have become the person I am now.

This thesis in particular could not have come about without the advice and support of Vahid Dabbaghian, Binay Bhattacharya, Patricia Brantingham, Richard Lockart, Andrew Park, Carl Schwartz, Fiona Young (all SFU) and Rob Thiessen of the Hope for Freedom Society.

Finally, and most importantly, I could not have done it without Denise and my family to fall back on.

[1] Members of a Chocolate-driven Social System

[2] see MoCSSy dictionary

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Homelessness is a paradox of the present time. It seems that, with global prosperity increasing, so too is homelessness. National governments are rising to the challenge and have set targets to reduce or even end homelessness, for example in England, The Netherlands, Scotland and the United States of America [17, 59, 14, 44, resp.]. Also the Canadian government has made homelessness a focal point of policy through its National Homelessness Initiative (NHI), announced on December 17, 1999 [18, 25]. Its goal was to reduce and prevent homelessness across Canada by helping communities address local homelessness issues through the creation of Sustaining Community Partnership Initiatives (SCPIs).

Greater Vancouver was one of the initial 61 communities with "demonstrated homelessness problems" [26] in which an SCPI was set up [5]. The Regional Steering Committee on Homelessness (RSCH), a key player in the SCPI for Greater Vancouver, set a more ambitious goal than its granting programme, the NHI, as it aimed to "eliminate homelessness in Greater Vancouver" [55]. Ten years later, despite the continued effort and support from the federal government [27], Greater Vancouver continues to face a serious homeless crisis: the results of regional homeless counts (conducted in 2002, 2005, 2008 and 2011) show a steady increase in the number of people who are homeless [31, 56, 11, 13]. The latest official estimate of the total number of homeless in Greater Vancouver is from the March 2008 homeless count: at that time, 2,660 homeless were counted. [56].

1.1 Changing 'Homelessness'

> "Homelessness is a broad term that can encompass a range of housing conditions. These can be understood on a continuum of types of shelter:
>
> - At one end *absolute homelessness* is a narrow concept that includes only those living on the streets or in emergency shelters.
> - *Hidden or concealed homelessness* is in the middle of the continuum. These include people without place of their own who live in a car, with family or friends, or in a long-term institution.
> - At the other end of the continuum, *relative homelessness* is a broad category that includes those who are housed but who reside in substandard shelter and/or who may be at risk of losing their homes" [15]

This thesis, and the data used in it, concerns itself with *absolute homelessness* only.

The interpretation of what it means to be 'homeless', and consequently the approaches taken to address homelessness, has changed markedly over the past 20 years [59]. Homelessness is now generally viewed as a social problem rather than just a personal one, a social perspective which is reflected well in the changing meaning of "home," from 'house' to 'place of belonging'. In my native language, Dutch, this shift has been even more explicit: 'dakloze' ('roof-less') became 'thuisloze' ('home-less') within my lifetime. This perspective presents a difficulty in developing programmes that aim to reduce homelessness: it is easier to describe the process of housing provision for an individual as an entity without context than it is to describe the process of a social individual finding a place of belonging. Yet the two are intricately linked: social support is an important factor in the long-term success of housing provision, and *vice versa* [59].

Social support is usually referred to as a 'social network,' and in the context of homelessness this buzz word [48] is used for a variety of other quite distinct concepts. In outreach and social work, a homeless person's social network consists of those people with whom this person had a positive relationship before becoming homeless, or those with whom (s)he might develop a relationship that would help him/her overcome homelessness [36, 59, 62]. The few cases in which relationships *between homeless* are studied refer to them as 'company' or 'community' rather than 'social network,' and conclude that these relationships are a significant factor in maintaining homelessness [21, 47]. In health and criminology, a

homeless person's social network consists of those other homeless with whom this person has a relationship [16, 32, 54, 67]. The effect of these relationships is sometimes positive, for example in reducing the incidence of depression [67], and sometimes negative, for example in encouraging deviant behaviour [32]. From these studies, no definite conclusion can be drawn about the net effect of social relations among homeless on homelessness.

Virtually all studies mentioned here approach social networks among homeless from a qualitative point of view. In this thesis, we take a quantitative point of view in studying data collected by the Hope for Freedom Society, a homeless outreach organisation that operates in the Tri Cities area (Coquitlam, Port Coquitlam and Port Moody) of the Greater Vancouver Metropolitan Area, British Columbia (B.C.), Canada. The Hope for Freedom Society is first and foremost a drug rehabilitation organisation that runs two recovery centres (Resurrection House and Glory House). Its outreach activities with on-the-street homeless started in 2006, when it was approached by representatives from the B.C. Ministry for Employment and Income Assistance (MEIA) to initiate a 6 month pilot project to establish a connection with the homeless population in the Tri Cities and assess what could be done to help homeless individuals overcome homelessness. After succesful completion of the initial pilot project the Hope for Freedom Society continued its work with funding from B.C. Housing, and the data collected in the following two years are the subject of this thesis.

We extracted information about social interaction from the data (chapter 2), and used these data to investigate social interaction among the Hope for Freedom Society's clients to shed light on some questions raised in earlier research, and by the data set itself:

- in focus group discussions with homeless drug addicts in the Vancouver Downtown East Side (DTES) [68], participants said drug preferences divided the community into several tightly-knit "cliques."
  - is this also the case in our study area (the Tri Cities)?
  - can we find strong sub-communities? (section 3)
  - if so, do they align with drug preference, or any other personal traits? (chapter 4) The Hope for Freedom Society suggests that that in the Tri Cities, there seems to be a tight-knit group of long-term homeless [66].
- reference [47] states that social relationships among homeless are a significant factor in maintaining homelessness. If this is so,

  - is the outreach organisation (unintentionally) strengthening these relationships by providing venues for interaction?
  - can we find evidence for positive social interactions? (section 5.2)

- can we extract actionable insights from the data, that the outreach organisation could use to improve their outreach?

The connecting thread of our analysis is the search for a principled way of determining the strength of a relationship between two individuals, based on interaction (chapter 3), personal characteristics (chapter 4) and personal activity spaces (chapter 5).

Chapter 2

Data Extraction

The data set used in this study was collected over a period of 2 years as part of the outreach activities of the Hope for Freedom Society in the Tri Cities area of B.C., Canada, an area consisting of three municipalities (Coquitlam, Port Coquitlam and Port Moody) with a total population of approximately $195,000$ [9]. The Hope for Freedom Society has approximately 120 homeless clients at any given point in time [64, 65, 66], and it maintains detailed logs of its activities and its clients. Each individual with whom the outreach workers are in sustained contact has a personal file in which the client's information is recorded. These personal profiles may contain:[1]

- name (and nickname);
- date of birth (or age);
- place of origin;
- appearance (sex, height, weight, eye colour, hair colour, tattoos and other features);
- government IDs (Social Insurance Number, Personal Health Number and Ministry of Employment and Income Assistance number);
- drug of choice;
- job skills;

[1] the presence and accuracy of information in a person's file depend on their ability and willingness to provide it.

- how long this person has been in the region;
- how long this person has been homeless; and
- a log of interactions with this person.

The Hope for Freedom Society uses this information to account for how it spends its funding, and to provide better services for its clients. Beyond the day-to-day provision of services, it also wants the information it collects to be used to further research into the factors that influence (aspects of) homelessness in the Tri Cities [64, 65], and has kindly provided access to their data for this thesis.

Our main focus in this thesis will be on the social network as captured by the Hope for Freedom Society's log of interactions with each person.[2] Each interaction log contains a wealth of information about the interaction between outreach workers and the person the file is associated with, but it also registers co-occurrence of, and interactions between, clients of the Hope for Freedom Society.

The first step in defining a network based on these data is to establish where to draw the boundary [34], determining what to include and what not to include. Each observation mentions individuals and locations, so these are candidates for inclusion. Not all individuals mentioned should be represented as vertices, however. For example, if an individual is reported to be convinced that he is John Lennon, that should not mean John Lennon is included in the network. Less hypothetically, it is unclear if mention of a client's sister should include her in the network. Two 'natural' boundaries may be conceived for the network: the Hope for Freedom Society's database forms an 'imposed' boundary on the network, delineating exactly those actors who are clients of the Hope for Freedom Society and about whom some personal information is known, but the logs contain approximately twice as many actors as are in the database, so the logs themselves could also be taken as the boundary of the network. We decide this trade-off between data quality and network (sample) size in favour of data quality, keeping only those actors for whom the Hope for Freedom Society has a client file. This limits our ability to investigate the effect of social support from non-homeless but allows us to incorporate personal information for every actor.

[2] From a methodological point of view, the information in these logs was collected by participant observers (the outreach workers) who follow a regular schedule of observations (presence at locations in the Tri Cities) supplemented by random observations (mostly personal service and emergency calls).

2.1 Privacy

These data are privacy-sensitive. Instead of stating this as a factoid, let us discuss the matter briefly, because privacy is a much misunderstood in the public domain [10]. The narrative of privacy often over-emphasises the need to restrict access to individual pieces of information, and neglects the importance of linkages between pieces of information (*i.e.* the semantic web in which they are embedded). For example, the number 301088019 is not privacy-sensitive on its own, nor is the number 922670095. Both these 9-digit numbers are numbers without any intrinsic meaning. When linked with the additional piece of information that the first is an SFU Student ID number, and the second a Canadian Social Insurance Number (SIN), the latter becomes potentially more privacy-sensitive than the former. This is only potentially so because although knowledge about which SFU ID and SIN numbers actually exist is not in the public domain, neither can be connected to any individual by the information given so far. A link between those two pieces information, *e.g.* that they belong to the same individual (they do not), *is* privacy-sensitive, *i.e.* more identifying than the information offered by both numbers separately: the first couple of digits of the SIN indicate that it concerns a non-resident, and the first couple of digits of the SFU ID indicate that this SFU student started their academic career at SFU after 2006. The connection reduces the possible individuals to which this information belongs to "international SFU students who started their academic career at SFU after 2006," and the SIN number possibly gives some additional indication about the date of entry into Canada (SIN issuance date).

Much of the personal information that the Hope for Freedom Society collects is far too privacy-sensitive to be looked at. Information on appearance and government IDs have been discarded immediately. Other, more useful information has been anonymised. Anonymisation in this context demands more than mere removal of personal identifiers; it demands that no conjunction of information about an individual would allow association of any file with an individual.

Two examples will make clear exactly how demanding this constraint is. Suppose there is only one person in the data set who is over 75 years of age. Clearly, the ages would have to be categorised to prevent this from being identifiable in the data. One could, for example, categorise age as <30, 30–50 and >50. Now suppose there was only one woman in the data set who is over 50 years of age. She would still be identifiable within the data set with just these two pieces of information. This can be carried through to all data fields in the set,

Variable ($N = 256$)	# missing	min	max	median	$\bar{X}$	s_X
Age (yrs)	17	8	73	42	40	9.8
Time Homeless (m)	40	1	216	8.5	16.5*	22*
Time in the Tri Cities(yrs)	172	1/12	45	3	7.4	10.3

$*$ two outliers (12 and 18 years) affect these statistics. Without these outliers, the mean would be 15 and the standard deviation 15.4.

Table 2.1: Descriptive statistics for continuous variables. None of the variables are normally distributed ($p \ll 10^{-10}$ Kolgomorov-Smirnov)

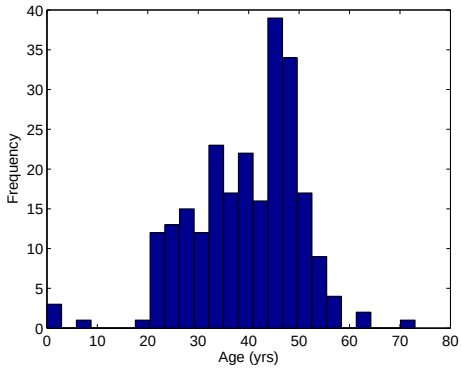
and the resulting requirement is then that no element of the refinement of all *observable* categorisations contains a 'small' group of individuals. Unfortunately, this refinement is subject to a combinatorial explosion, which is extremely problematic when the number of records is small, as is the case for this data set ($N = 256$).

The data fields that could (and should) be so categorised are:

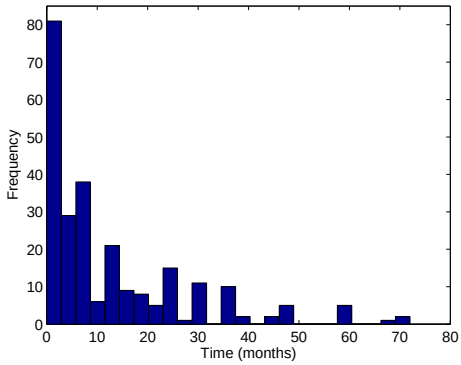
1. Sex,
2. Drug of Choice,
3. Job Skills,
4. Place of Origin,
5. Age,
6. how long this person has been in the region (Time in the Tri Cities), and
7. how long this person has been homeless (Length of Homelessness).

Reducing each of these to just a binary (Sex and Job Skills already are) would still produce 128 refined categories, with in expectation 2 individuals per category. Information provided by the Hope for Freedom Society suggests that 3 and 4 may not be relevant for our purposes. Furthermore, 6 is missing in 67.6% of the cases. At risk of discarding valuable information, but in the interest of being able to work with this data at all, we only retain the following categorised variables:

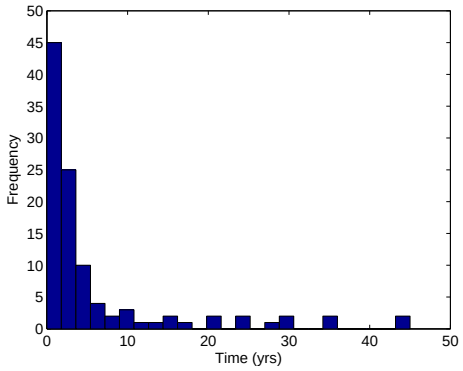
- Sex (male, female, unknown)
- Drug of Choice (alcohol, heroine, cocaine, crystal meth, pot, unknown)



(a) Age



(b) Length of Homelessness (2 outliers omitted: 12 and 18 years)



(c) Time in the Tri Cities

Figure 2.1: Histogram plots of continuous variables

- Age (<33 years, 33–40 years, 40–47 years, >47 years, unknown)
- Length of Homelessness (<6 months, 6–15 months, >15 months, unknown)

The small number of observations in the Drug of Choice category "pot" (see 2.2) suggests that it needs to be categorised as "unknown". It is worth noting that both Length of Homelessness and Drug of Choice are only identifying within this data set, as neither are observable or recorded elsewhere. Moreover, an actor's drug of choice records a preference. Actual drug use is much less clearly categorised, to the point that most clients of the Hope for Freedom Society are generalist poly-drug users [66]. Drug of Choice is retained, however, because it may be indicative of other, unrecorded and/or unobservable information about clients, such as their social behaviour (*e.g.* cocaine addicts may be more socially isolated due to drug-related paranoia [57, 66]). In the end, we are left with 8 observable categories, the refinement of Age and Sex.

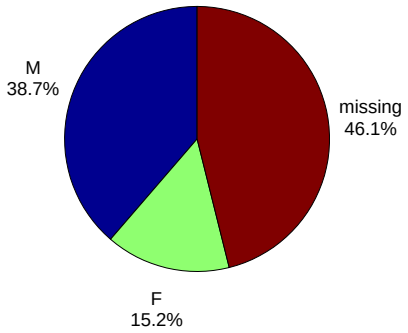
Each of the data fields mentioned on page 8 was extracted from the data separately, and permuted randomly to destroy association with the original data set and other variables to compute the descriptive statistics listed in 2.1 (continuous variables; see also 2.1) and 2.2 (categorical and categorised variables; see also 2.2). For continuous variables, standard descriptive statistics are reported; for categorical variables, frequency counts as well as statistics on the extracted social interactions are reported. These statistics on extracted social interactions will be discussed in 2.2.1.

Interaction Logs

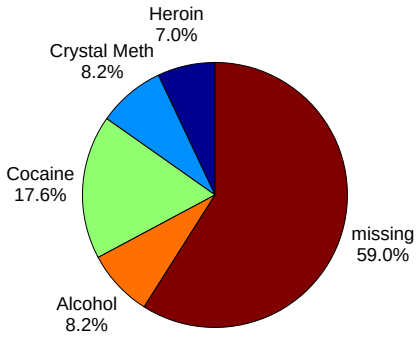
With the anonymisation of personal information done, we now turn to the more challenging task of anonymising the raw text data of the interaction logs. We could not read these logs due to privacy constraints, so an alternative approach was required. The actual textual information is not important for our purposes, and neither is it for anonymisation. It suffices that those words and other strings that identify individuals or locations be marked and replaced by a random but consistent identifier. This can be done by looking at each word separately, one at a time. The words in the logs have meaning only conditional upon other words, *i.e.* actual words are privacy-sensitive only in context, as argued on page 7: the word "hoegaerdt" by itself, for example, is not associated with anyone. It may or may not be a name, and without the cue of capitalisation it is hard to tell.

Overall	N		N_{obs}		N^*_{obs}	N^*_{obs}/N_{obs}
	256	(100%)	11,105	(100%)	10,082	90.79%
Sex	N		N_{obs}		N^*_{obs}	N^*_{obs}/N_{obs}
Female	39	(15.23%)	2097	(18.88%)	1936	92.32%
Male	99	(38.67%)	5465	(49.21%)	4992	91.34%
Unknown	118	(46.09%)	3543	(31.90%)	3514	89.02%
Drug of Choice	N		N_{obs}		N^*_{obs}	N^*_{obs}/N_{obs}
Alcohol	21	(8.20%)	936	(8.43%)	745	79.59%
Cocaine	45	(17.58%)	3013	(27.13%)	2910	96.58%
Crystal Meth	21	(8.20%)	1687	(15.19%)	1458	86.43%
Heroine	18	(7.04%)	786	(7.08%)	716	91.09%
Pot	6	(2.34%)				
Unknown	145	(56.64%)	4683	(42.17%)	4253	90.82%
Job Skills	N		N_{obs}		N^*_{obs}	N^*_{obs}/N_{obs}
Has skills	63	(24.61%)				
Does not have skills	193	(75.39%)				
Age	N		N_{obs}		N^*_{obs}	N^*_{obs}/N_{obs}
< 33 years	57	(22.27%)	2237	(20.14%)	2146	95.93%
33–40 years	53	(20.70%)	3014	(27.14%)	2764	91.71%
40–47 years	62	(24.22%)	2330	(20.98%)	1793	76.95%
> 47 years	67	(26.17%)	3197	(28.79%)	3071	96.06%
Unknown	17	(6.64%)	327	(2.94%)	308	94.19%
Length of Homelessness	N		N_{obs}		N^*_{obs}	N^*_{obs}/N_{obs}
< 6 months	73	(28.52%)	1912	(17.22%)	1741	91.06%
6–15 months	83	(32.42%)	3106	(27.97%)	2697	86.83%
> 15 months	60	(23.44%)	4621	(41.61%)	4216	91.24%
Unknown	40	(15.63%)	1466	(13.20%)	1428	97.41%
Time in the Tri Cities	N		N_{obs}		N^*_{obs}	N^*_{obs}/N_{obs}
< 2 years	24	(9.38%)				
≥ 2 years	60	(23.44%)				
Unknown	172	(67.19%)				

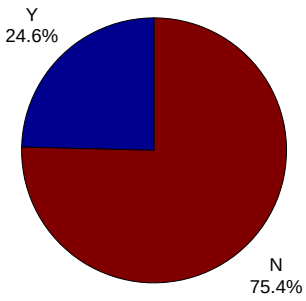
Table 2.2: Frequency counts for categorical and categorised personal attributes, with total number of observations (N_{obs}), of which most are unambiguous (N^*_{obs}; see 2.2.1). N^*_{obs} counts the number of times we could unambiguously assign an observation of a name, for example, to an actor who prefers cocaine.



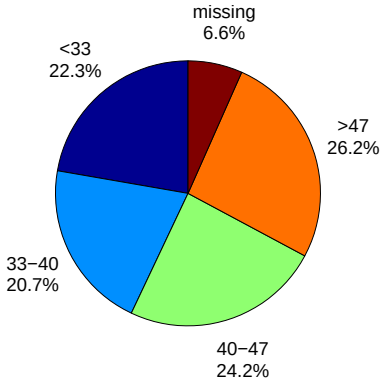
(a) Sex



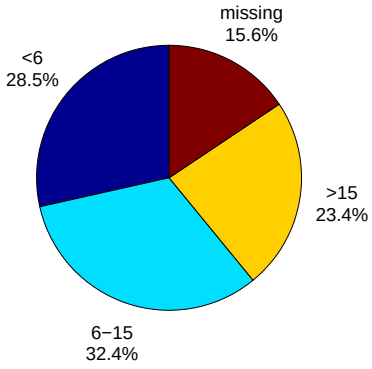
(b) Drug of Choice



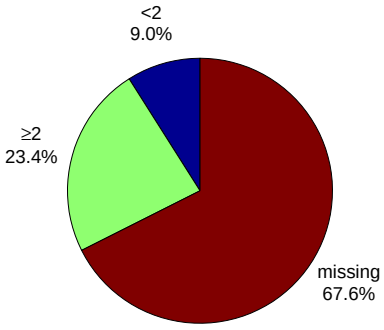
(c) Job Skills



(d) Age (categorised)



(e) Length of Homelessness (categorised)



(f) Time in the Tri Cities (categorised)

Figure 2.2: Categorical variables and categorised continuous variables

In this way, we anonymised these data without looking at the actual text, but by instead extracting a list of individual unique words (6283 in all, as opposed to 110,242 in the original text), sorted alphabetically to break syntactic relations. The words were stripped of numbers, apostrophes, punctuation, capitalisation and any other semantics beyond the word level. We then categorised unique words as either names (privacy-sensitive), locations or irrelevant words (not privacy-sensitive). Similar names were grouped and replaced in the text by a representative from the group they belonged to, for example "John," "jhon" and "Johnm" were all replaced by "john". Finally, the list of names was used to hash occurrences of these names in the logs to unique random numerical values.

We note that this approach may well scale because the number of unique words in a text increases logarithmically with the total length of the text [6, 38, 50].

2.2 Social Network

After anonymisation, and with the word lists (locations, names, other) in hand, the logs could readily be converted into hypergraphs representing several (social) networks.

Before proceeding, let us introduce some notation and definitions. Readers who are unfamiliar with social network analysis may refer to Appendix B for an introduction into the use of hypergraphs for social network analysis. See also reference [63] for a similar conceptual model.

Definition 2.2.1. A hypergraph $\mathcal{H}(V, \mathcal{E})$ consists of a set of n vertices V and a set of m hyperedges $\mathcal{E}$. A vertex $v_i \in V$ represent an entity, *e.g.* an actor (client), a name or a location. A hyperedge $\varepsilon_i \in \mathcal{E}$, $\varepsilon_i \subset V$ represents a connection, *e.g.* an interaction (observation) or a profile, between one or more entities, and may occur multiple times (*e.g.* the same group of people observed at different dates).

Because we will be using several different hypergraphs in our analyses, denote ${}_x\mathcal{H}^y$ the hypergraph with x as edges and y as vertices, *e.g.* ${}_{obs}\mathcal{H}^{loc}$ is a hypergraph of observations (as hyperedges) containing location names (as vertices) and ${}_{prof}\mathcal{H}^{name}$ is a hypergraph of client profiles (as hyperedges) containing names (as vertices). This notation derives from the computational representation of a hypergraph as an incidence matrix $\mathbf{I}$:

$$\mathbf{I} = [\mathbf{i}]_{m\times n}, \quad \mathbf{i}_{ij} = \begin{cases} \textsc{True} & \text{if } v_j \in \varepsilon_i \\ \textsc{False} & \text{otherwise} \end{cases} \tag{2.1}$$

in which each row represents a hyperedge, and each column represents a vertex, so it is the function table of the relation "vertex j belongs to edge i". As with $\mathcal{H}$, let ${}_x\mathbf{I}^y$ denote the incidence matrix of the hypergraph with x as edges and y as vertices.

From a hypergraph one can derive two 'projections,' graphs that represent bilateral association between vertices or edges.

Definition 2.2.2. A graph $G(V, E)$ that is a vertex projection of a hypergraph $\mathcal{H}(V, \mathcal{E})$ consists of the set of vertices V of $\mathcal{H}$ and a set of edges E. Two vertices are connected by an edge $e_{ij} = e_{ji} = \{v_i, v_j\} \in E$ in G if they are connected by a hyperedge in $\mathcal{H}$.

A hyperedge projection $G(\mathcal{E}, E)$ is similar: its 'vertices' are the hyperedges of the hypergraph, and two hyperedges are connected if they share a vertex.

Similar to $\mathcal{H}$, let ${}_xG^y$ denote the graph with y as vertices and sharing of x as edges, *e.g.* ${}_{loc}G^{obs}$ is a graph of observations connected when they mention the same location. It is the hyperedge projection of ${}_{obs}\mathcal{H}^{loc}$.

A graph is represented computationally as an $n \times n$ adjacency matrix $\mathbf{A}$. Elements $\mathbf{a}_{ij}$ in the adjacency matrix of the vertex projection of a hypergraph $\mathcal{H}(V, \mathcal{E})$ count the number of hyperedges (ε) in which the two vertices v_i and v_j occur together

$$\mathbf{A} = [\mathbf{a}]_{n\times n}\,,\ \mathbf{a}_{ij} = \left|\left\{\varepsilon \in \mathcal{E} \;\middle|\; v_i, v_j \in \varepsilon\right\}\right| \tag{2.2}$$

Similarly, for the hyperedge projection, elements $\mathbf{a}_{ij}$ count the number of vertices the two hyperedges ε_i and ε_j share:

$$\mathbf{A} = [\mathbf{a}]_{m\times m}\,,\ \mathbf{a}_{ij} = \left|\varepsilon_i \cap \varepsilon_j\right|$$

As with G, let ${}_x\mathbf{A}^y$ denote the adjacency matrix of the graph with y as vertices and sharing of x as edges. The relationship between ${}_x\mathbf{A}^y$ and ${}_x\mathbf{I}^y$, and ${}_y\mathbf{A}^x$ and ${}_x\mathbf{I}^y$ is simple (using TRUE = 1 and FALSE = 0):

$${}_x\mathbf{A}^y = \left({}_x\mathbf{I}^y\right)^{\mathsf{T}}\,{}_x\mathbf{I}^y \tag{2.3a}$$

$${}_y\mathbf{A}^x = {}_x\mathbf{I}^y\left({}_x\mathbf{I}^y\right)^{\mathsf{T}} \tag{2.3b}$$

2.2.1 Network Extraction

We will use several hypergraphs to discuss separate aspects of the data: names mentioned in observations (${}_{obs}\mathcal{H}^{name}$), locations mentioned in observations (${}_{obs}\mathcal{H}^{loc}$) and names mentioned

in personal profiles (${}_{prof}\mathcal{H}^{name}$) (see also Table 2.3). These, and the dates of observation and profile in the log of which the observation was recorded, could be extracted straightforwardly by a one-pass scan of the observations, because observations are well delimited in the interaction log.

An example (hypothetical) interaction log could look like

> (Ali) 01/17/07 Met Peter and Mary near Coquitlam River. Petres bene clean now for two weeeks, he says. His sister Shaema has been pushin ghim to go for recovery.
> { Kees] 10/05/06 At PoCo City Hall, gave coffee and snacks.

All observations follow the same pattern: they start with some descriptive information, a date and the name of an outreach worker in brackets. Therefore, a one-pass scan of these observations can correctly determine the boundaries between observations and identify (using the word lists) client names, staff names, the observation date and locations:

> (⟨⟨N24⟩⟩) 01/17/07 Met ⟨⟨N184⟩⟩ and ⟨⟨N80⟩⟩ near ⟨⟨L136⟩⟩ ⟨⟨L24⟩⟩. ⟨⟨N184⟩⟩ bene clean now for two weeeks, he says. His sister ⟨⟨N444⟩⟩ has been pushin ghim to go for recovery.
> { ⟨⟨N4⟩⟩] 10/05/06 At ⟨⟨L2⟩⟩ ⟨⟨L96⟩⟩ Hall, gave coffee and snacks.

Note that our approach to anonymisation and relevant word identification is not affected by mild misspellings (provided diligence in anonymising), but prohibits identification of words that are locational only in context (*e.g.* "Hall"). Staff and client names were hashed by the same process, as clients and staff may have the same name (they do not), but are stored separately to retain as much information as can be extracted.

It is crucial that the aforementioned hypergraphs accurately represent the data, so that they can form a solid basis for the construction of two additional hypergraphs, associating observations with profiles (${}_{obs}\mathcal{H}^{prof}$; the 'true' social network) and locations with profiles (${}_{loc}\mathcal{H}^{prof}$). We now turn to the construction of ${}_{obs}\mathcal{H}^{prof}$, which will then allow us to construct ${}_{loc}\mathcal{H}^{prof}$ by linking the association between observations and profiles to the association between observations and locations:

$$ {}_{loc}\mathbf{I}^{prof} = {}_{obs}\mathbf{I}^{loc^{\mathsf{T}}}\,{}_{obs}\mathbf{I}^{prof}, $$

so that a location name is connected to a profile if that location has been visited by the person the profile is associated with (that person and location were mentioned together in one observation).

Notation	Hyperedges	Vertices	semantics
${}_{obs}\mathcal{H}^{name}$	Observations	Names	each observation joins names of clients who are related in some way
${}_{obs}\mathcal{H}^{loc}$	Observations	Locations	observations may mention one (or sometimes multiple) locations
${}_{prof}\mathcal{H}^{name}$	Profiles	Names	each profile contains a first name, a last name (20 missing) and possibly a nickname (in 38 cases)
${}_{obs}\mathcal{H}^{prof}$	Observations	Profiles	the actual social network: co-observation of clients
${}_{prof}\mathcal{H}^{loc}$	Profiles	Locations	a geographical network of clients at locations

Table 2.3: Hypergraphs extracted (upper 3) and derived (lower 2) from the data

Construction of ${}_{obs}\mathcal{H}^{prof}$ In linking observations to profiles (constructing ${}_{obs}\mathcal{H}^{prof}$) we made use of all of the already determined links between observations and profiles. For example, the second observation of the hypothetical example on the preceding page does not have any client name in it, but it is clearly referring to the client in whose profile the observation was recorded, and will be linked to the file in the log of which it occurred. Conversely, this reasoning was used to remove from the observation those names associated with the profiles it was thus unambiguously linked to ("Peter" in the hypothetical example on the previous page), resolving 0.5% of connections between names and observations. Next, unique first and last names were used to assign observations to the profiles these names belonged to, resolving most connections (70.9%). Names that did not appear in any profile (individuals who were not in sustained contact with the Hope for Freedom Society; see on page 5) were removed from the observations altogether (16.43%). There was no overlap between the outreach workers' and clients' names.

After these matchings, 40 names and 12.2% of connections in ${}_{obs}\mathcal{H}^{name}$ were left unresolved (not assigned to a person's profile). These were cases in which multiple people (profiles) had the same name and the association was not resolved by the previous matching. In 2.2 on page 11, we have broken down the total number of observations (N_{obs}) and the number of unambiguous observations of individuals (N^{*}_{obs}) by personal characteristics. Drug of Choice category 'Alcohol' and Age category '40–47 years' stand out as having lower

percentages of unambiguous observations, which is likely due to two individuals with many ambiguous observations (see also Figure 2.3). Note that Table 2.2 was constructed *after all observations had been resolved* in order to make the division by personal attributes, and that the overall percentage of ambiguous observations of people (N^*_{obs}/N_{obs}) is therefore different from the 12.2% reported here, which is for ambiguous observations of *names*.

We formulated the remaining disambiguation task (*e.g.* "given several observations of a 'John,' which ones refer to John Doe, and which to John Smith?") as a clustering problem: every observation is a point in space, with coordinates $\{0, 1\}$ representing the absence or presence of a name or person. The clusters in this many-dimensional space represent individuals. In applying this reasoning, we assumed that the group of people (other names) an individual (name) occurs with can help determine who they are. This reasoning is deeply embedded in folklore, for example in the hispanic proverb "Dime con quién andas, y te diré quién eres" [Tell me who you walk with, and I will tell you who you are], and we can actually verify it, too.

Most observations of persons with an ambiguous name are not ambiguous. A name may occur in multiple profiles, and therefore be an ambiguous name, but there are usually many more unambiguous observations of the people this name may be associated with (see Figure 2.3). On average, 76.3% of observations are unambiguous, providing a substantial ground truth to validate the clustering against.

We used this ground truth to select an appropriate clustering algorithm from those available in MATLAB: Cluster-Data and K-Means. These algorithms represent two conceptually different approaches to clustering: optimising boundary consistency (points are close to some other point(s) in their cluster) or internal consistency (points are close to the centre of their cluster), respectively. Each comes with a host of distance functions[3] to make precise what 'close' means, and we compared all of them using our ground truth. Cluster-Data consistently outperformed k-Means, although only by a narrow margin (see Figure 2.4).

From among the distance measures, we selected the best, Pearson's correlation coefficient [45, 53] to parameterise Cluster-Data with, so if X_1 and X_2 are observations (points

[3]Cluster-Data: Chebychev distance, cityblock (taxicab) distance, Cosine distance, euclidean distance, Hamming distance, Jaccard coefficient, Mahalanobis distance, Minkowski distance, Pearson's correlation, Spearman rank correlation, and standardised euclidean distance
K-Means: cityblock (taxicab) distance, Cosine distance, Hamming distance, Pearson's correlation and standardised euclidean distance

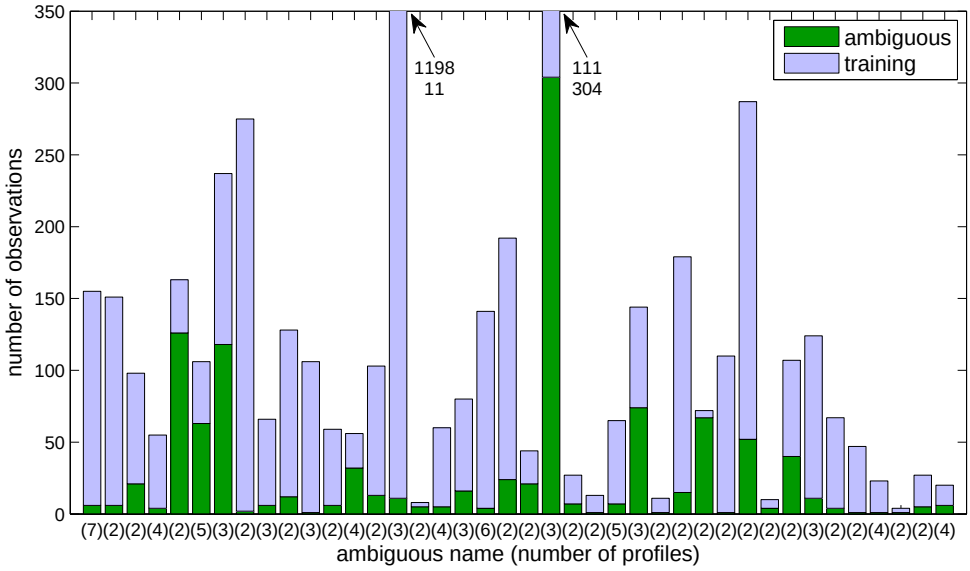


Figure 2.3: Ambiguous observations of names *v.* unambiguous observations of the persons this name could be associated with (used as training data and ground truth on the previous page). Each bar represents one ambiguous name (*e.g.* 'John'), with (in brackets) the number of profiles containing this name. The lower bar represents the number of times this name ('John') is observed but cannot be unambiguously assigned to a profile. The upper bar represents the number of unambiguous observations of all individuals with this name (Johns) in the data.

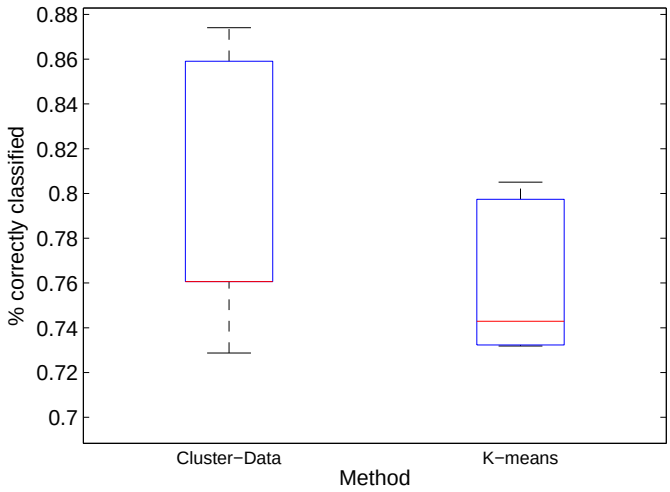


Figure 2.4: Comparison of the two clustering algorithm available in MATLAB: CLUSTER-DATA and K-MEANS (average performance per distance measure)

in many-dimensional space), the distance between them is

$$r_\delta(X_1, X_2) = 1 - \frac{X_1 - \bar{X}_1}{\sqrt{SS_{X_1}}} \cdot \frac{X_2 - \bar{X}_2}{\sqrt{SS_{X_2}}}, \quad SS_X = (X - \bar{X}) \cdot (X - \bar{X}) \tag{2.4}$$

where $X \cdot X = X^\mathsf{T} X$ denotes the dot product of X and itself, and the subscript δ indicates that the formula has been adjusted to measure *distance* rather than *similarity*.

2.3 Descriptive Statistics

Only now it is possible to calculate descriptive statistics for the network. Some results were already tabulated in Table 2.2, showing a tendency for individuals with missing personal information to be observed less frequently. The causation is likely reversed: personal information in clients' files is updated as outreach workers interact with them, and fewer interactions simply provide less opportunity for completing the files. This apparent tendency may, however, be the result of mere chance. An Analysis of Variance (ANOVA) failed to reject the null hypothesis that groups were similar ($p < 0.05$) for all personal attributes but Length of Homelessness, where short-term homeless (< 6 months) are observed less frequently ($p < 0.05$) than long-term homeless (> 15 months; using Scheffé's correction [58] for

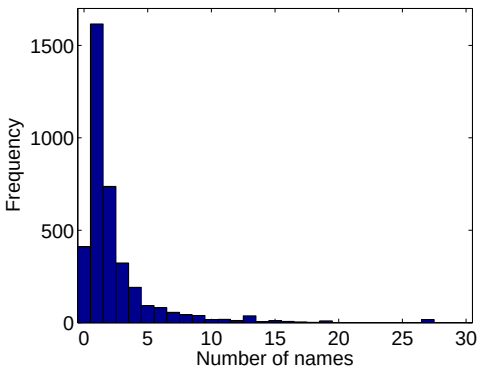
comparisons of multiple means), and the category 'missing' was not significantly different from any other groups. The significant difference for Length of Homelessness is expected: short-term homeless may not have been present since the Hope for Freedom Society started recording (see also section 5 and Figure 5.1).

Global distributions of observation frequency and observation 'size' (in number of people or number of locations) are reported in Figure 2.5. Several outliers (specified in the figure captions) were omitted from the histogram plots to improve visual resolution. These outliers in the number of observations mentioning a location name may be the result of misclassification in the anonymisation process. It is possible that a commonly occurring word was mistakenly tagged as a location when it was not. Therefore, these location names were removed from the data and will not be considered in further analyses. The outliers in the number of observations containing a person's name are not the result of misclassification: both outlier names occur in a client's profile (and only in that one profile). Consequently, mentions of these names in observations are correctly interpreted as observations of the two clients.

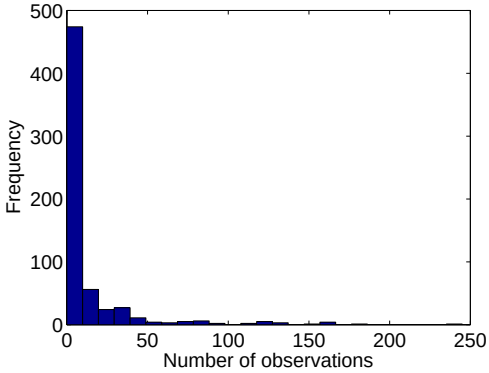
One more aspect of the data should be illuminated here: time (see Figure 2.6). The observation density (*i.e.* number of observations per month) increases over time, as should be expected with data going back to the very first month in which the Hope for Freedom Society was active. The increase in observation size is likely in part due to the starting up of some programmes in which larger groups of clients participate at the same time. It may, however also be a result of data entry fatigue: several observations may be recorded as one to avoid having to log multiple. The apparent dip in observation frequency in the summer of 2008 could be seasonal but additional years of data would be required to establish if this effect is structural. It likely is, since demand for services is quite weather-dependent [66], and the Hope for Freedom Society also runs a number of winter-only programmes (*e.g.* the Cold and Wet Weather Mat Programme [4]).

2.3.1 Network Properties

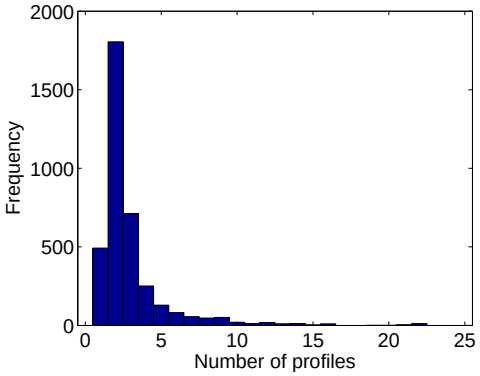
As described in 2.2 on page 14, several simple graphs can be derived from the hypergraphs extracted here. Rather than treating all possible derived graphs, let us focus on the interactions between people, the core of this thesis, and between locations, which will be of use for chapter 5. In 2.5 on page 27 we list several statistics for four derived networks (see Table 2.4). Below we briefly explain the statistics, for which we will use some additional



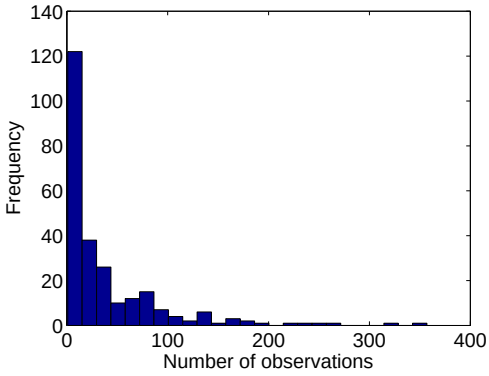
(a) Number of names per observation



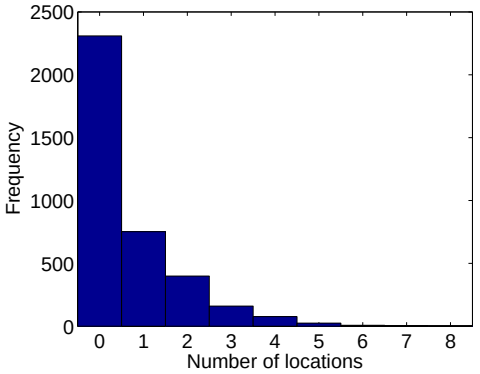
(b) Number of observations per name (2 outliers omitted: 304, 1166)



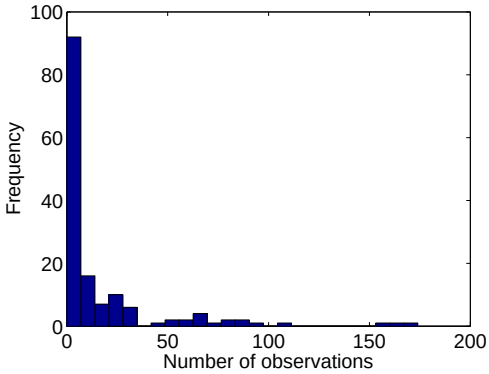
(c) Number of profiles per observation



(d) Number of observations per profile (1 outlier omitted: 1204))



(e) Number of location names per observation



(f) Number of observations per location name (3 outliers **removed**: 1121, 1196, 1490)

Figure 2.5: Relation distributions (edge and vertex degrees) in the extracted hypergraphs

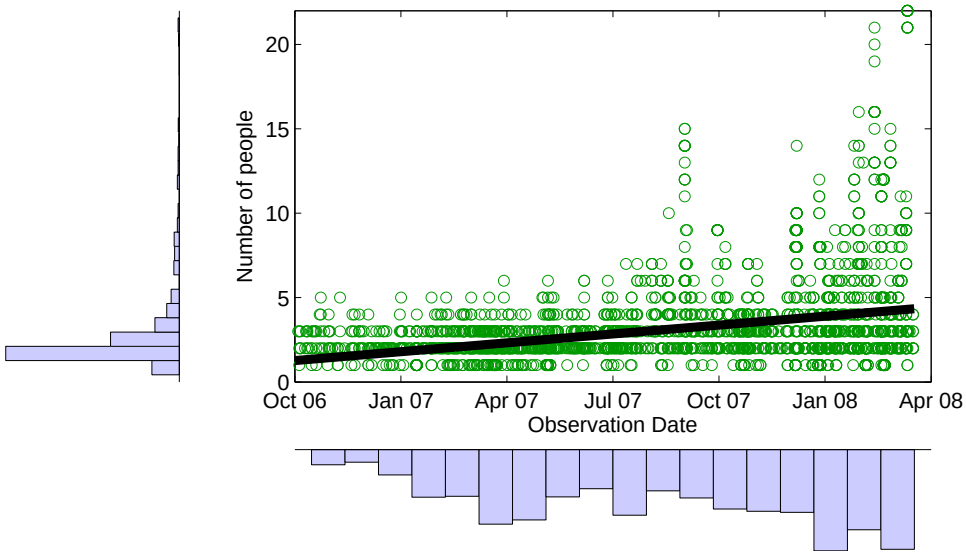


Figure 2.6: Number of people mentioned in an observation over time, with fitted linear regression line ($R^2 = 0.104,\ p < 10^{-10}$; see Equation 4.4)

Notation	A =	Vertices	semantics
${}_{obs}G^{prof}$	${}_{obs}\mathbf{I}^{prof\,\mathsf{T}}\,{}_{obs}\mathbf{I}^{prof}$	Profiles	'the' social (interaction) network, in which edges represent co-observation
${}_{loc}G^{prof}$	${}_{prof}\mathbf{I}^{loc\,\mathsf{T}}\,{}_{prof}\mathbf{I}^{loc}$	Profiles	individuals connected by an edge visited the same location, possibly at different times
${}_{obs}G^{loc}$	${}_{obs}\mathbf{I}^{loc\,\mathsf{T}}\,{}_{obs}\mathbf{I}^{loc}$	Locations	a semantic web of location names, connecting location names that were mentioned in the same observation
${}_{prof}G^{loc}$	${}_{prof}\mathbf{I}^{loc}\,{}_{prof}\mathbf{I}^{loc\,\mathsf{T}}$	Locations	an insight into the activity space of the Hope for Freedom Society's clients, connecting locations if they were visited by the same person.

Table 2.4: Selected derived graphs, with semantics (using ${}_{prof}\mathbf{I}^{loc} = {}_{obs}\mathbf{I}^{prof\,\mathsf{T}}\,{}_{obs}\mathbf{I}^{loc}$; See Table 2.2.1)

notation that will help write formulae concisely.

Definition 2.3.1. In a graph $G(V, E)$, the neighbourhood N_i of a vertex v_i consists of those vertices v_j that v_i is connected to by an edge:

$$N_i = \left\{ v_j \in V \;\middle|\; \{v_i, v_j\} \in E \right\}$$

The size of v_i's neighbourhood is its degree $d(v_i)$.

Density The density (ρ) of a graph is a measure of how 'full' or densely connected a graph is.

$$\rho = \frac{m}{\binom{n}{2}} \tag{2.5}$$

Typically, social networks have a low density.

Distance

Distance in a graph is measured over paths between vertices.

Definition 2.3.2. A path p, denoted $v_i \rightsquigarrow v_j$, $(v_i, v_j \in V)$, is a sequence of edges $p = \left(\{v_i, v_{k_1}\}, \{v_{k_1}, v_{k_2}\}, \ldots \{v_{k_\ell}, v_j\}\right)$ connecting two vertices that does not visit any vertex more than once. The length of a path $\ell(p)$ counts the number of edges on the path. The distance $\ell(i,j)$ between two vertices v_i and v_j is the length of the shortest path between these two vertices:

$$\ell(i,j) = \min_{p = v_i \rightsquigarrow v_j} \ell(p)$$

The distribution of path lengths in a graph is usually summarised using the distribution of largest shortest path lengths from each vertex:

$$\ell_k = \left| \left\{ v_i \;\middle|\; \max_j \ell(i,j) = k \right\} \right| \tag{2.6}$$

Perhaps due to the importance of this distribution in graph theory, its minimum and maximum have been given special names: radius (rad) and diameter ($\varnothing$), respectively. These are also reported in Table 2.5, together with the average $\bar{\ell}$

Connected Component Using the notion of paths, one can identify connected components in a graph as those groups of vertices in which each vertex can reach each other via a path. The number of vertices in the largest such component (n_{cc}) is reported in Table 2.5. In ${}_{obs}G^{prof}$, our main object of analysis, the largest connected component is also the only connected component, and the other vertices are isolated (see Figure 2.7).

Transitivity: Clustering Coefficient (γ)

The clustering coefficient or transitivity of a graph measures how much a person's friends are also friends among themselves. The local clustering coefficient for a vertex v_i measures the 'realised potential' of connections among neighbours of v_i as the proportion of possible connections $\left(\binom{d(v_i)}{2}\right)$ that actually exist.

$$\gamma_i = \frac{\left|\{e = \{v_j, v_k\} \mid e \in E, v_j, v_k \in N_i\}\right|}{\binom{d(v_i)}{2}} \tag{2.7}$$

Note that the similarity to Equation 2.5; the local clustering coefficient measures the density of links among the neighbours of v_i.

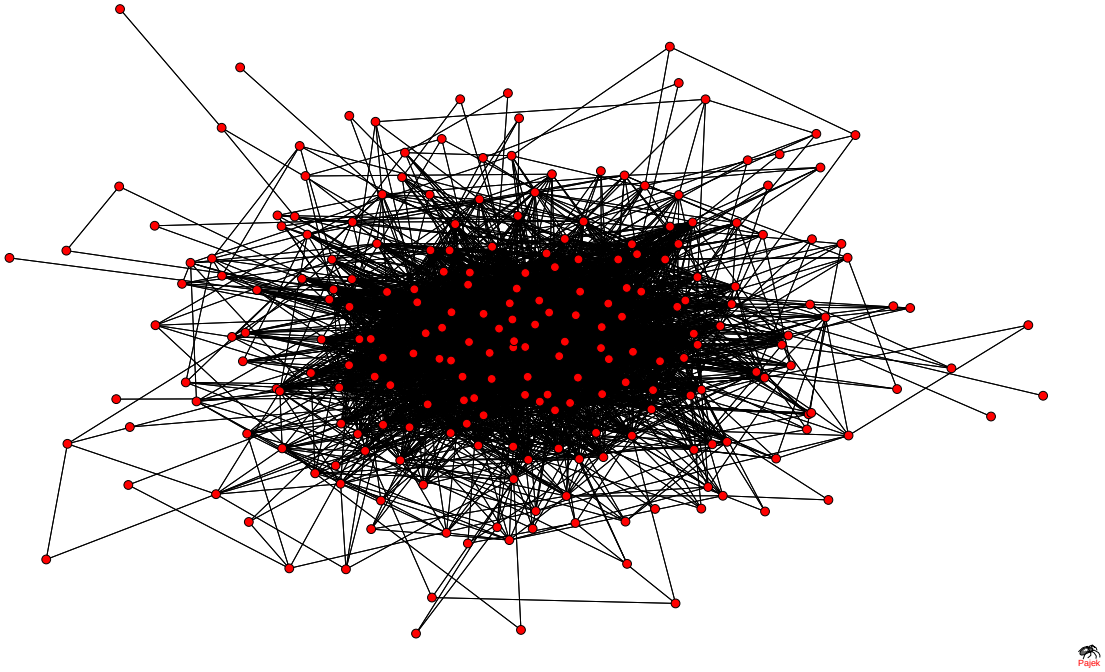


Figure 2.7: ${}_{obs}G^{prof}$ (20 isolated vertices omitted), diagrammed using the Kawada-Kamai spring mass graph layout algorithm [33]

The global clustering coefficient, the average local clustering coefficient over all vertices, gives an indication of how tightly knit communities in the graph are:

$$\gamma = \frac{3 \times \text{number of triangles}}{\text{number of 2-paths}} \tag{2.8}$$

Centrality & Centralisation

Centrality is a measure of how important (central) an actor is in the network, and can be defined in many ways. Centrality measures are usually used to rank vertices in the graph, and analyse an intervention (*e.g.* removal) targetting the most central actors [20, 35, 63], but they can also be used to describe balance in the network by computing the corresponding centralisation, a single-number measure of the disparity between centrality of vertices [34]. The centrality of a vertex v_i by centrality measure x is written as $C_x(i)$. The corresponding centralisation C_x is then the average disparity with respect to the largest centrality of any vertex:

$$C_x = \max_j \big(C_x(j)\big) - \frac{\sum_i C_x(i)}{n} \tag{2.9}$$

A high centralisation means that the distribution of vertex centralities is skewed, which is an indication that there may be a few very central actors in the network.

A number of centrality measures have been defined [19]. We will use three of the most common ones: degree, closeness and betweenness.

Degree Centrality assumes that the degree of an actor reflects their activity within the network, and that more active actors are more central in the network:

$$C_d(i) = \frac{d\,(v_i)}{n-1} \tag{2.10}$$

This scales the degree to fall within $[0, 1]$, which is necessary for being able to compare centralisations.

Closeness Centrality reflects how close an actor is to other actors in the network. It is the inverse of the average distance from a vertex to all other vertices in the network.

$$C_{\ell^{-1}}(i) = \left(\frac{\sum_{j \neq i} \ell(i,j)}{n-1} \right)^{-1} \tag{2.11}$$

	n	n_{cc}	m	ρ	rad	$\bar{\ell}$	$\varnothing$	γ	$C_{\ell^{-1}}$	C_B	C_d
${}_{obs}G^{prof}$	256	236	3375	0.103	3	4.059	6	0.495	0.315	0.397	0.465
${}_{loc}G^{prof}$	256	224	19172	0.587	2	2.170	3	0.865	0.196	0.112	0.282
${}_{obs}G^{loc}$	150	129	663	0.059	3	3.775	5	0.340	0.247	0.216	0.266
${}_{prof}G^{loc}$	150	144	5057	0.453	2	2.076	3	0.703	0.256	0.336	0.473

Table 2.5: Statistics for selected derived networks

Actors that can 'reach' other actors in the network quickly are more central in the network, assuming every increment in distance costs more time.

Betweenness Centrality reflects how much of a hub for shortest paths an actor is. It is the proportion of all shortest paths that go through a vertex:

$$C_B(i) = \frac{\left|\left\{p = v_j \rightsquigarrow v_k \;\middle|\; \ell(p) = \ell(j,k), v_i \in p\right\}\right|}{\left|\left\{p = v_j \rightsquigarrow v_k \;\middle|\; \ell(p) = \ell(j,k)\right\}\right|} \tag{2.12}$$

where $\ell(p) = \ell(j,k)$ says that "p is a shortest path between v_i and v_j," Actors that are on shortest paths between many pairs of actors could moderate information that flows along those paths, and may therefore be influential in the network.

Application to ${}_{obs}G^{prof}$

The co-observation network (${}_{obs}G^{prof}$) is not very dense, as would be expected from a social network. The largest distance between any two vertices is 6, in keeping with the popular belief that in a social network, any pair of people is separated by at most six degrees [69]. This is higher than one would expect from a graph with uniformly random connections (rad = $\bar{\ell}$ = $\varnothing$ = 3), but this can be explained in conjunction with the relatively high clustering coefficient, and may be evidence of a core-periphery structure: as more connections are local within the neighbourhoud of a vertex, fewer edges are available to make the long-distance (non-local) connections that make paths short. All centralisations are quite high, indicating that there may be a few very central individuals, or a core group of individuals. Betweenness centralisation being slightly higher than closeness centralisation could be a point in favour of a few very central individuals, but this may be a consequence of betweenness being 0 for quite a few vertices (closeness never is for vertices in a connected component).

Application to ${}_{loc}G^{prof}$

The co-location network (${}_{loc}G^{prof}$) encodes information about the overlap between the activity spaces of actors, and this overlap is large indeed (58.7%). Still there is some evidence of localisation, found in the large clustering coefficient, much larger than would be expected from a graph with uniformly random connections ($\gamma \approx 0.587$). This large clustering coefficient is likely in part a consequency of how ${}_{loc}G^{prof}$ is constructed, connecting every pair of actors who visit a particular location, effectively creating a large cluster for every location (see Figure 3.1 on page 34).

Application to ${}_{obs}G^{loc}$

The location association network (${}_{obs}G^{loc}$) is a semantic web of location names: locations that are mentioned together in an observation are somehow related. This network is the most sparse of the networks treated here, but still has a high clustering coefficient, suggesting that some semantic grouping of locations may be possible. The high clustering coefficient may in part be an artefact of the way ${}_{obs}\mathcal{H}^{loc}$ was constructed. For example, in the hypothetical example given on page 15, "Poco City Hall" was converted to two separate locations: "⟨⟨L2⟩⟩ ⟨⟨L96⟩⟩ Hall." Three or more location names connected in this way may artificially increase the clustering coefficient somewhat, but this does not harm the interpretation: these location names are indeed highly semantically related. The distance distribution suggests that there may be a few peripheral locations ($\ell = 5$) that are only mentioned in special circumstances.

Application to ${}_{prof}G^{loc}$

The actor activity spaces, of which the joint structure is represented by ${}_{prof}G^{loc}$, are well-connected. The low average distance indicates that information about events at one location can spread geographically quite easily. With these small distances, it is not surprising that the closeness centralisation is not very high. The higher betweenness and degree centralisations suggest that there may be some geographical hubs that many actors frequent. It may also reveal that a few locations are frequently visited by outreach workers, increasing the number of observations at this location and thereby sampling more of the population at this particular location.

Chapter 3

Interaction-Based Edge Weights

We are now in a position to investigate social interaction among the Hope for Freedom Society's clients. We would particularly like to identify groupings and important individuals.

- Finding communities within the population would allow the development of group-specific strategies, and finding distinct parts within the joint activity space may help find effective locations to do outreach. High clustering coefficients (γ) for the location-based graphs suggest there is at least some local grouping, although they may also be artifacts of high density (ρ).

- Important individuals might be targets for specific action, or could be utilised to spread information among the population. The relatively high betweenness and closeness centralisations (C_B, $C_{\ell^{-1}}$) in the co-observation network suggest that indeed there may be such information brokers in the network.

Communities

Many community finding strategies have been developed (see reference [46] for a review), but few have been developed on as generic a probabilistic footing as those based on modularity [41, 42]. Modularity itself is not an algorithm but a measure of how 'community-ish' a division of a graph into communities is. The quality of such a community division is computed by comparing to a random null model, and subtracting the expected from the observed edge weight $\mathbf{w}$:

$$b_{ij} = \mathbf{w}(i,j) - E\left(\mathbf{w}(i,j) \,\middle|\, G\right) \tag{3.1}$$

Positive residuals (b_{ij}) in the modularity matrix B indicate that something social is going on between v_i and v_j. Intuitively, a 'good' community is one within which much social is going on, *i.e.* in which many pairs of vertices have positive b_{ij}. Modularity operationalises this by summing over all within-community b_{ij}. It is customary to use $b_{ii} = 0$, to avoid isolated vertices being good communities.

The most common null model used for modularity is a random graph in which the probability (and expected value) of an edge between two vertices is proportional to the degrees $d(\cdot)$ of these two vertices:

$$E\left(\mathbf{w}(i,j) \,\middle|\, G\right) = \frac{d\left(v_i\right) d\left(v_j\right)}{2m} \tag{3.2}$$

where m is the number of edges in the graph.

Now that we have the two building blocks, and using Kronecker's delta

$$\delta(i,j) = \begin{cases} 1 & v_i \text{ and } v_j \text{ are in the same community} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

to codify the communities, we can write modularity as [43]

$$q = \frac{1}{2m} \sum_{i,j} \left[\mathbf{w}(i,j) - \frac{d\left(v_i\right) d\left(v_j\right)}{2m} \right] \delta(i,j), \tag{3.4}$$

where a conventional $1/2m$ factor and

$$m = \frac{1}{2} \sum_{i'} d\left(v_{i'}\right)$$
$$d\left(v_i\right) = \sum_{j'} \mathbf{w}\left(i,j'\right)$$

scale q to fall within $[0,1]$. In this formulation communities may overlap, but most analyses and algorithms (ours also) restrict themselves to finding disjoint communities.

Modularity being a generic quality measure, it can be heuristically maximised (actual maximisation is NP-hard [7]) by a variety of general as well as special-purpose algorithms. The best algorithm to our knowledge was proposed by Mark Newman [43]. Although no approximation bound for this algorithm is known, in practice it performs well [43], and for the only benchmark for which a comparison with optimal modularity could be made (Zachary's karate network [71]), it is within 5% of the optimal.[1]

[1] optimal modularity for this network is reported in reference [7] and the performance of Newman's method is reported in reference [43].

Newman's algorithm was used to heuristically maximise modularity for ${}_{obs}G^{prof}$, ${}_{loc}G^{prof}$, ${}_{obs}G^{loc}$ and ${}_{prof}G^{loc}$, but the results are mostly disappointing: $q \approx 0.06$ for ${}_{loc}G^{prof}$ and ${}_{prof}G^{loc}$ (no community structure), and $q = 0.209$ for ${}_{obs}G^{prof}$ where $0.4 \leq q \leq 0.7$ is usual for most social networks [7, 12, 43]. Only ${}_{obs}G^{loc}$ shows some community structure ($q = 0.379$ for 7 communities of sizes $\{5, 13, 20, 20, 25, 27, 32\}$), indicating that locations can be divided into 7 categories such that locations from a particular category are usually mentioned in an observation with other locations from that category, and not with locations from other categories. These categories could represent functional domains (*e.g.* locations related to rehabilitation, health care, housing, *etc.*) or perhaps the activity spaces of each outreach worker (there are six). The result is interesting regardless of the underlying mechanism because although such categories can be constructed, these categories are not found in ${}_{loc}G^{prof}$ or ${}_{prof}G^{loc}$, indicating that whatever be the defining features of these categories, clients visit all of them. The evidence of structure found here will be used in chapter 5 to incorporate location information into an edge weighting scheme.

Information Brokers in ${}_{obs}G^{prof}$

Broadcasting information about a change in service hours or a new outreach initiative without proper targeted broadcast media can be a tedious and time-consuming task for the Hope for Freedom Society. This time could be just as well spent actually helping clients. The question that naturally arises is:

> "is it possible to let the clients spread the news themselves? And if so, who should the news be spread to?"

In large social networks, such questions would be answered approximately by selecting actors with high centrality (*e.g.* closeness), but since ${}_{obs}G^{prof}$ is quite small, it is possible to determine exactly who to spread the news to so that it reaches everyone, and to keep the number of people to contact as low as possible. This can be done by modelling the question as a set cover problem in which the sets are the immediate neighbourhoods (N_i) of the actors:

Definition 3.0.3. Given a set V a family of subsets $S = \{N_i \mid N_i \subset V\}$, find the smallest possible collection of subsets $S' \subseteq S$ such that their union equals V:

$$\min_{S' \subseteq S} \left|S'\right| \text{ such that } \bigcup_{N_i \in S'} N_i = V$$

Note that we restrict our analysis to the connected component of ${}_{obs}G^{prof}$, since the remaining vertices are all singletons, and would have to be contacted individually.

We solved the corresponding Integer Linear Programme to find the set cover represented by column vector c, in which TRUE values indicate that a vertex is in the set cover.

$$\min_{c} f^{\mathsf{T}} c \text{ such that } \begin{cases} {}_{obs}\mathbf{A}^{prof} c \geq b \\ c \text{ integer} \end{cases} \tag{3.5}$$

where

$$f = [1]_{n\times 1} \qquad\qquad b = [1]_{n\times 1}$$

using MATLAB's BINT-PROG function. Despite the NP-hardness of the set cover problem, an optimal solution was obtained in 0.613 seconds on a 2008 laptop computer.[2] This fast solution (for a random graph of same size and density, the ILP takes multiple days to solve) suggests that the graph contains quite some structure that was exploited by the ILP solver, and it does. For example, consider following the greedy algorithm

- find vertices of degree 1, and select their single neighbour for the set cover;
- remove all vertices now covered (the neighbours of all vertices in the set cover); repeat.

When applied to ${}_{obs}G^{prof}$, it cuts down the size of the graph to only 12 vertices before it can no longer find vertices of degree 1. The efficacy of this strategy is further evidence for a core-periphery structure as hypothesised on page 27.

The smallest possible group that can reach all clients (set cover c) is rather large: it contains $|c| = 27$ actors, approximately 11.4% of clients (n_{cc}) in the connected component. This may be a disadvantage because we were looking for an efficient way to reach clients, and the overhead of specifically targetting these 27 actors may outweigh the gain from letting the network do the information spreading. On the other hand, on average a vertex in the set cover reaches 9 vertices not reached by others, but the average degree is 28.6 for the whole graph, and 31.3 for the vertices in the connected component. Because of this, 70% of vertices are covered more than once. This overlap is good for spreading information, because individuals are then not dependent on one provider of information.

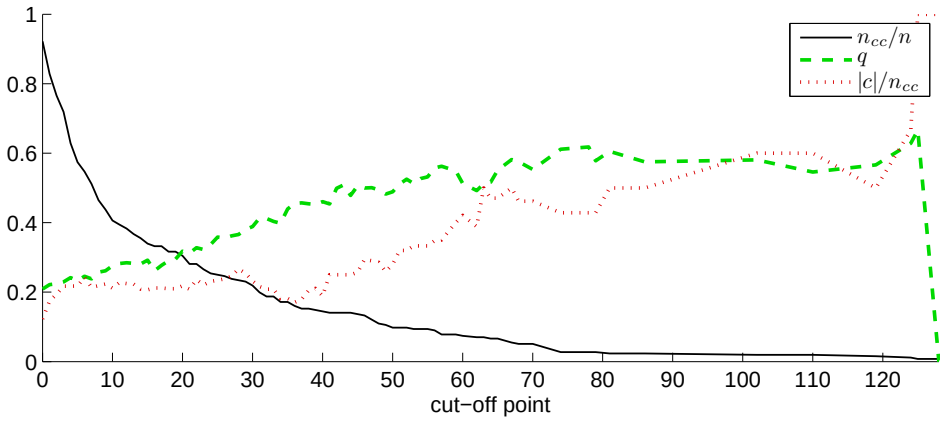


Figure 3.1: Proportion of vertices in the largest connected component (n_{cc}/n), modularity (q) and proportion of the vertices in the connected component that are in the set cover ($|c|/n_{cc}$) for increasing co-observation frequency cut-off. Only edges with frequencies larger than the cut-off point are retained

Cutting the Low-Frequency Interactions

Perhaps the 'real' results are better (higher modularity, smaller set cover) than presented in the previous section. It is well possible that the results are influenced by inaccuracies in the data in the form of spurious co-observations of individuals who do not actually know each other. Since these inaccuracies likely originates from misrecordings, random encounters recorded in the data, and errors in anonymising and disambiguating the data, most of it should occur as low-frequency interactions. Cutting out these low-frequency interactions may improve the network and also focus the results on the more active and possibly more essential relationships in the network. The effect this cutting has on the network size, modularity and the size of the set cover is shown in Figure 3.1. To give an impression of what this means for the actual network, Figure 3.2 shows snapshots of the network with a selection of different thresholds for what is 'low-frequency' and therefore cut from the network.

To the extent that communities are present in the network, the increase in modularity achieved by Newman's method [43] as low-strength edges are cut from the network indicates that edges connecting communities are being cut more than edges within communities, in

[2] 2 GHz dual core processor, 2GB RAM

line with Mark Granovetter's famous observation/hypothesis that weak ties connect intimately connected communities [22, 23]. However, 80% of actors are disconnected from the network (Subfigure 3.2(f)) before modularity reaches a reasonable level of modularity. The relative size of the set cover actually increases to around 25%, making information spread even more robust to failure of transmission, but also more time-consuming to initiate.

These snapshots do not seem to change much as more edges (and vertices) are cut out, supporting the existence of a nested core-periphery structure, as suggested by the analysis of descriptive statistics on on page 27 and corroborated by the set cover experiment on page 31. The fact that this structure is persistent even when when 60% of the vertices in the network have been cut out (Subfigure 3.2(e)) shows that the core is not only more connected, but also connected by higher-frequency interactions.

Though not directly useful due to the rapid shrinking of the connected component, these results do suggest that co-observation frequency counts have a meaningful role to play in the network, and it is to this role that we now turn.

3.1 Social Influence as Link Strength

The construction of $\mathbf{A}$ as $\mathbf{I}^{\mathsf{T}}\mathbf{I}$ results in a matrix in which elements $\mathbf{a}_{ij} \in \mathbf{A}$ represent absolute frequencies of co-observation:

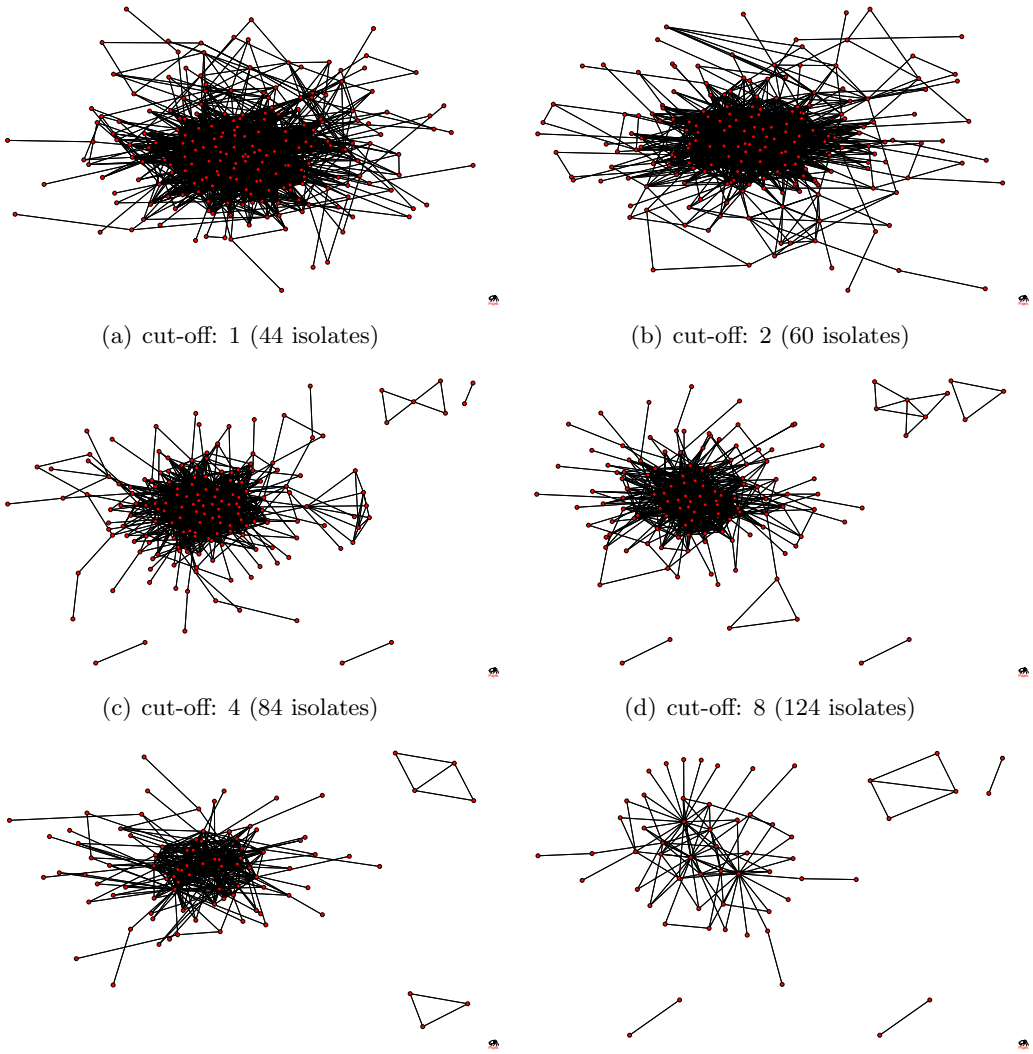
$$\mathbf{w}(i,j) = \sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj} \tag{3.6}$$

In other words, observations of size x are replaced by complete graphs of size x, and the resulting edge weights between vertices are summed over all observations to give a co-observation frequency (see Figure 3.3 for an example). This may be useful in itself, but we would rather attempt to compute a weighting $\mathbf{w}(i,j)$ between vertices v_i and v_j that conveys information about less concrete but more powerful concepts such as *influence* of one actor over another.

Column normalisation [34] is a quite common way of deducing information about influence from a social network that has weights associated with the edges. It uses the degree of the 'target' vertex j, which reference [34] defines as

$$d\left(v_j\right) = \sum_{i \neq j} \mathbf{a}_{ij} \qquad \left(\text{usually } \mathbf{a}_{jj} = 0\right) \tag{3.7}$$

to scale the weights so that the resulting edge weights represent each actor's proportional



(a) cut-off: 1 (44 isolates)

(b) cut-off: 2 (60 isolates)

(c) cut-off: 4 (84 isolates)

(d) cut-off: 8 (124 isolates)

(e) cut-off: 16 (164 isolates)

(f) cut-off: 32 (198 isolates)

Figure 3.2: ${}_{obs}G^{prof}$ for various co-observation frequency cut-off points (isolated vertices omitted), diagrammed using the Kawada-Kamai spring mass graph layout algorithm [33] in Pajek [3]. See Figure 2.7 for the original (uncut) network.

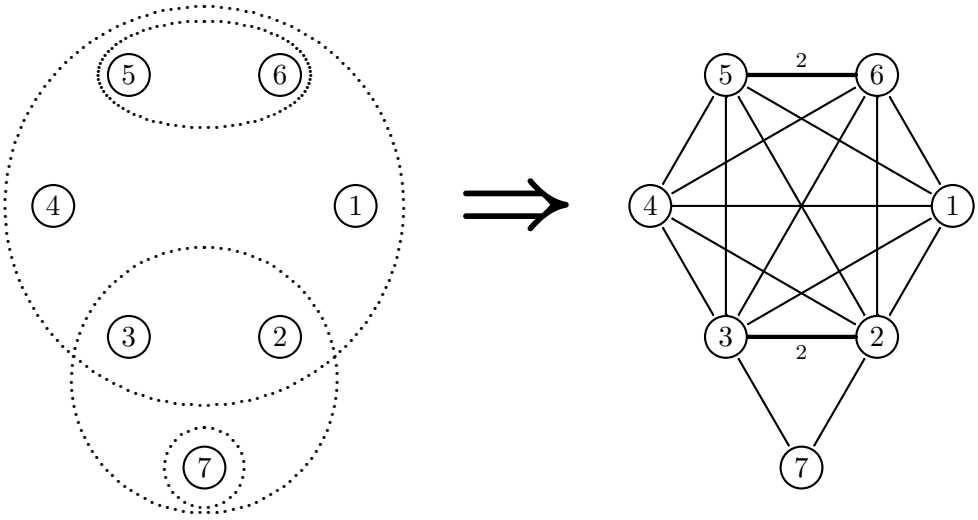


Figure 3.3: The problem of co-observation frequency as connection strength: when $\mathbf{A} = \mathbf{I}^{\mathsf{T}}\mathbf{I}$, a complete graph is substituted for every hyperedge

participation in an interaction with another actor.

$$\mathbf{w}(i,j) = \frac{\mathbf{a}_{ij}}{d\left(v_j\right)}, \tag{3.8}$$

For example, suppose elements $\mathbf{a}_{ij}$ in the adjacency matrix count the number of letters sent from v_i to v_j. A column-normalised weight $\mathbf{w}(i,j) = 0.41$ then means that of all letters sent to v_j, 41% came from v_i. Row normalisation works analogously, scaling by the degree of the 'source' vertex i.

In the context of our data set, elements $\mathbf{a}_{ij}$ in the adjacency matrix count the number of co-observations of v_i and v_j. Applying column normalisation directly to the example of Figure 3.3 would, for example, give $\mathbf{w}(2,3) = 2/6$, supposedly meaning that of all observations of v_3, 33% are with v_2. Clearly, it should be 100%. This naïve application does not accurately reflect the conceptual underpinning of column normalisation, because the degree as defined in Equation 3.7 is not an accurate representation of total activity.

In this chapter, rather than just correcting the calculation by using the degree as defined in 2.3.1 on page 23, we take the intuition behind column normalisation several steps further, and develop a more elaborate edge weighting scheme on that footing, because weighting schemes are usually defined on $\mathbf{A}$, which, being a projection of $\mathbf{I}$, contains less information than is available. The adjusted operationalisations of the edge weighting schemes were designed to make use of this extra information while remaining true to the original concept.

After introducing our extension to the underlying intuition (section 3.2) and making it precise (subsection 3.2.1), we guide the reader through a series of edge weighting schemes that incorporate ever more of the intuition sketched in section 3.2. At every step along the way, we check our progress against criteria set out in subsection 3.2.1.

3.2 Intuition

The weight $\mathbf{w}(i, j)$ of edges in the graph represented by $\mathbf{A}$ should measure *influence*, operationalised as a neighbour v_i's *relative importance* to the current vertex v_j. To make precise *relative importance* we assume that an individual who wants to exert influence needs to dedicate effort (time) to doing so, and that an individual will be influenced more by people who invest more effort in influencing him/her. From this follow several desirable properties:

1. the influence of a neighbour v_i over the current vertex v_j increases
   - as they interact more, and
   - as they have stronger interactions (see property 3);

   i.e. more effort invested translates to more influence.
2. influence depends on the total influence received (sent), *i.e.* it takes more to (be) influence(d by) a vertex that has many contacts (high degree) than one that has few (low degree); and
3. the probability (and strength) of an interaction within a meeting goes down as more people are involved in a meeting.

Each of these are explained below.

3.2.1 Operationalisation

Property 1 translates into monotonicity of weight with respect to the number of co-observations. We could operationalise property 1 as

$$\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj} < \sum_k \mathbf{i}_{kp} \cdot \mathbf{i}_{kq} \leftrightarrow \mathbf{w}(i, j) < \mathbf{w}(p, q), \tag{3.9}$$

so that if v_i and v_j are observed together less frequently than v_p and v_q, our weighting scheme should give the edge between v_i and v_j a lower weight than the edge between v_p and

v_q. The implication is bidirectional because the left hand side is symmetric with respect to i, j, p and q. A consequence of this definition which will be useful in later proofs is that

$$\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj} = \sum_k \mathbf{i}_{kp} \cdot \mathbf{i}_{kq} \leftrightarrow \mathbf{w}(i,j) = \mathbf{w}(p,q), \tag{3.10}$$

but this means that $\mathbf{w}(i,j)$ can only be an element-wise monotonic function on $\mathbf{A}$.

Consequently, Equation 3.9 is not useful for our purpose, as we expressly wish to find a weighting function that conveys something more than what is already captured in $\mathbf{A}$. The following list describes one alternative weaker operationalisation of Equation 3.9 and three possible subsets of Equation 3.9 that may hold for functions that are not an element-wise mapping of $\mathbf{A}$. These will constitute the criteria that will be used to guide the development of weighting schemes in this chapter.

(*adding an interaction*) (Strict) Monotonicity of $\mathbf{w}(i,j)$ with respect to co-observation: the weight $\mathbf{w}(i,j)$ should monotonically increase with co-observation of v_i and v_j, *i.e.* adding one co-observation of v_i and v_j should (*ceteris paribus*) not decrease $\mathbf{w}(i,j)$. Strict monotonicity is desirable for $\mathbf{w}(i,j)$, $i \neq j$ but not necessarily for $\mathbf{w}(i,i)$, *e.g.* in the case of column normalisation on $\mathbf{A}$ (see Figure 3.1 on page 34), in which all elements in a column i are divided by $\mathbf{w}(i,i)$ to give proportional co-observation ("$\mathbf{w}(i,j)$% of the time v_j was observed, (s)he was observed with v_i").

If $\mathbf{i}_{k\ell}$ and $\mathbf{w}(i,j)$ represent the original interactions and edge weights, and $\mathbf{i}'_{k\ell}$ and $\mathbf{w}(i,j)'$ the interactions and edge weights after a single interaction between v_i and v_j has been added, the above translates to

$$\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj} \leq \sum_k \mathbf{i}'_{ki} \cdot \mathbf{i}'_{kj} \leftrightarrow \mathbf{w}(i,j) \leq \mathbf{w}(i,j)'$$

This requirement **(*adding an interaction*)** should be maintained at all times. The following should be broken, since they 'measure' how close a weighting scheme still is to being an element-wise monotonic function of $\mathbf{A}$. Each is here presented with a positive interpretation, to illustrate the reasons for inclusion in this listing.

(*diagonal comparison*) Strict Monotonicity along the diagonal

if $\mathbf{w}(i,i)$ is strictly monotonic in the sense of the **(*adding an interaction*)**, it is also possible that it is strictly monotonic with activity when compared to other diagonal elements $\mathbf{w}(j,j)$, *i.e.* that

Equation 3.9 $\mid i = j,\ p = q$

This would allow the diagonal elements to be interpreted as a measure of overall activity, as in $\mathbf{A}$, in which the diagonal contains the total number of observations (degree) of the actors.

(*diagonal = max*) Relationship between $\mathbf{w}(i,j)$, and $\mathbf{w}(i,i)$ or $\mathbf{w}(j,j)$:

since both $\mathbf{w}(i,i)$ and $\mathbf{w}(i,j)$ are measures of activity (vertex and between-vertex, respectively), they should use the same currency (units). Hence, it may be desirable that the diagonal is the maximal element in its row and/or column:

Equation 3.9 | $i = p = q$

Intuitively, an actor's influence over him-/herself should be at least as much as that of any other actor over him/her, since at least as much time is spent with oneself as with others.

(*within-row/-column comparison*) Within-row and -column monotonicity:

by extension of the previous, it may be desirable that an element $\mathbf{w}(i,j)$ can not only be compared to the diagonal elements of the row and column it is in, but also to the other elements in its row or column:

Equation 3.9 | $i = p$

3.2.2 Inverse Proportionality

Property 1 only required the contribution of a single co-observation k to $\mathbf{w}(i,j)$ to depend on $\mathbf{i}_{ki} \cdot \mathbf{i}_{kj}$, but in order to capture property 2 and 3, it should also depend (inversely) on the number of interactions v_j has, as well as the number of other actors involved in each interaction between v_i and v_j:

$$\mathbf{w}(i,j) = \sum_k g\left(\mathbf{i}_{ki} \cdot \mathbf{i}_{kj},\ \left(\sum_\ell \mathbf{i}_{\ell i}\right)^{-1},\ \left(\sum_\ell \mathbf{i}_{\ell j}\right)^{-1},\ \left(\sum_\ell \mathbf{i}_{k\ell}\right)^{-1}\right) \tag{3.11}$$

where g is a function that will be developed throughout this chapter.

To avoid extraneous notation, use

$$d(v_i) = \sum_\ell \mathbf{i}_{\ell i}$$

$$d(\varepsilon_k) = \sum_\ell \mathbf{i}_{k\ell}$$

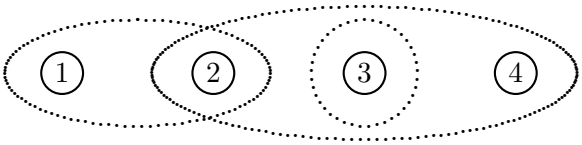


Figure 3.4: Counterexample used throughout section 3.3

wherever the details of the sum are unimportant, so that Equation 3.11 is

$$\mathbf{w}(i,j) = \sum_k g\left(\mathbf{i}_{ki} \cdot \mathbf{i}_{kj},\ d(v_i)^{-1},\ d\left(v_j\right)^{-1},\ d(\varepsilon_k)^{-1}\right) \tag{3.12}$$

Throughout we will use 3 parameters to fine-tune the weighting scheme: β_{out}, β_{in} and β_ε, which are associated with $d(v_i)$, $d\left(v_j\right)$ and $d(\varepsilon_k)$, respectively.

3.3 Candidate Edge Weighting Schemes

In this section we introduce a sequence of edge weighting schemes by concept, common operationalisation, and their desirability as per section 3.2.

Throughout we will mostly use the same counterexample (see Figure 3.4) to show that requirements from subsection 3.2.1 are broken:

$$\mathbf{I} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad \text{(counterexample)}$$

3.3.1 Column Normalisation on I

As our first step in extending the intuition behind column normalisation on $\mathbf{A}$ we investigate column normalisation on $\mathbf{I}$, taking the conceptual motivation from the previous subsection and apply it (meaningfully) to $\mathbf{I}$ by weighting an interaction by the inverse of the total number of interactions of an actor:

$$\frac{\mathbf{i}_{ki}}{d(v_i)}.$$

Thereby satisfying property 2. The interpretation is still the same: $\mathbf{w}(i,j) = 0.41$ means that of all interaction that v_j (v_i) is involved in, 41% is with v_i (v_j). The only difference is now that $\sum_{k \neq j} \mathbf{w}(k,j)$ no longer needs to sum to 100% because actors may be observed on their own (*e.g.* vertex 3 in Figure 3.4).

The resulting weights are then

$$\mathbf{w}(i,j) = \sum_k \left(\frac{\mathbf{i}_{ki}}{d(v_i)^{\beta_{out}}} \cdot \frac{\mathbf{i}_{kj}}{d(v_j)^{\beta_{in}}} \right) \tag{3.13}$$

Desirability A look at the diagonal elements

$$\mathbf{w}(i,i) = \sum_k \frac{\mathbf{i}_{ki}^2}{d(v_i)^{\beta_{out}+\beta_{in}}} = d(v_i)^{1-(\beta_{out}+\beta_{in})}$$

makes clear that it would be desirable to restrict the values of the parameters such that $\beta_{out} + \beta_{in} \leq 1$ to maintain positive monotonicity for (***adding an interaction***), which is strict if the inequality is.

(***adding an interaction***) SATISFIED

Since $\mathbf{i}_{ij} \in \{\text{TRUE}, \text{FALSE}\}$ and we are only adding one interaction between v_i and v_j (*ceteris paribus*), when we compare the numerator and denominator, after adding an interaction ($\mathbf{i}'$) over before ($\mathbf{i}$), it is clear that the numerator grows more rapidly than the denominator. This is because the interaction added may not increase $\sum_\ell \mathbf{i}_{\ell j}$, but it must always increase $\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj}$.

$$\frac{\sum_k \mathbf{i}'_{ki} \cdot \mathbf{i}'_{kj}}{\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj}} \geq \frac{\sum_\ell \mathbf{i}'_{\ell j}}{\sum_\ell \mathbf{i}_{\ell j}}$$

Therefore, $\mathbf{w}(i,j)$ is indeed monotonic with $\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj}$, and strictly positively so if $\beta_{out} + \beta_{in} < 1$.

(***diagonal comparison***) SATISFIED

$$\mathbf{w}(i,i) = \frac{\sum_k \mathbf{i}_{ki}^2}{d(v_i)^{\beta_{out}+\beta_{in}}} = d(v_i)^{1-(\beta_{out}+\beta_{in})}$$

diagonal elements can be compared as measures of activity if $\beta_{out} + \beta_{in} < 1$.

(***diagonal = max***) SATISFIED (Column) NOT SATISFIED (Row)

Diagonal elements are the maximum of their column

$$\frac{\sum_k \mathbf{i}_{ki}}{\cancel{d(v_i)^{\beta_{out}} \cdot d(v_j)^{\beta_{in}}}} \geq \frac{\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj}}{\cancel{d(v_i)^{\beta_{out}} \cdot d(v_j)^{\beta_{in}}}} \leq \frac{\sum_k \mathbf{i}_{kj}}{\cancel{d(v_i)^{\beta_{out}} \cdot d(v_j)^{\beta_{in}}}}$$

but not of their rows:

Counterexample. (using (counterexample) and Equation 3.10)

$$\sum_k \mathbf{i}_{k1} = \sum_k \mathbf{i}_{k1} \cdot \mathbf{i}_{k2}$$
$$\mathbf{w}(1,1) = 1 \neq \mathbf{w}(1,2) = \frac{1}{2^{\beta_{in}}}$$

↯

(*within-row/-column comparison*) NOT SATISFIED

The different denominators across rows and columns break the within-column monotonicity

Counterexample. (using (counterexample) and Equation 3.10)

$$\sum_k \mathbf{i}_{k2} \cdot \mathbf{i}_{k1} = \sum_k \mathbf{i}_{k2} \cdot \mathbf{i}_{k3}$$
$$\mathbf{w}(1,2) = \frac{1}{2^{\beta_{in}}} \neq \mathbf{w}(3,2) = \frac{1}{2^{\beta_{out}+\beta_{in}}}$$

and similarly for within-row monotonicity. ↯

This first edge weighting scheme has already shown a few general properties of an edge weighting scheme that will fulfill the requirements set in subsection 3.2.1:

- β_{out} and β_{in} need to be upper bounded in order to keep **(*adding an interaction*)**;
- the **(*diagonal comparison*)** condition cannot be broken by β_{out} or β_{in}; and
- **(*diagonal = max*)** is a special case of **(*within-row/-column comparison*)**.

Now with some more understanding of normalisation by vertex degrees, let us turn to the role of interaction size.

3.3.2 Row Normalisation on I

Row normalisation on $\mathbf{I}$ is not merely analogous to column normalisation: it is special because the effects of the normalisation and the dot product's summation ($\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj}$) are orthogonal to one another, thereby incorporating information from all of $\mathbf{I}$ into each $\mathbf{w}(i,j)$.

In our data, the rows of $\mathbf{I}$ are co-observations of actors, and the rationale for normalisation is that larger meetings have weaker interactions (see subsection 3.2.2):

$$\frac{\mathbf{i}_{ij}}{d(\varepsilon_k)}$$

The reasoning behind this is that on average, when two actors v_i and v_j are involved in two interactions ε_{k_1} and ε_{k_2}, and $|\varepsilon_{k_1}| < |\varepsilon_{k_2}|$, we would expect that the chance of an actual interaction between v_i and v_j occurring in ε_{k_1} be larger than the same chance in ε_{k_2}, and also (in part consequently) that the expected strength (duration) of the interaction in ε_{k_1} be larger than in ε_{k_2} (*ceteris paribus*).

The resulting weights are then

$$\mathbf{w}(i,j) = \sum_k \left(\frac{\mathbf{i}_{ki}}{d(\varepsilon_k)^{\beta_\varepsilon}} \cdot \frac{\mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}} \right) \tag{3.14}$$

satisfying property 3 on page 37.

Desirability No restriction needs to be put on β_ε from the point of view of maintaining or breaking any of the properties described in subsection 3.2.1, but a requirement based on interpretability should be made. As mentioned, any pair of individuals should be less likely to have a strong actual interaction as the size of the interaction (hyperedge) increases, but only to a limited extent: the total amount of interacting going on should intuitively still increase with increasing interaction size, and should at least not decrease for any individual.[3] Therefore, we require that $2\beta_\varepsilon \leq 1$.

(*adding an interaction*) SATISFIED

Observe that $\frac{\mathbf{i}_{ki}}{d(\varepsilon_k)}$ is constant for all i and strictly postive. Every TRUE in $\mathbf{I}$, then, is replaced with some strictly positive value, which is enough to satisfy the monotonicity condition with respect to adding an interaction.

(*diagonal comparison*) NOT SATISFIED

Diagonal elements can still be compared as activity under our changed definition that takes into account the interaction sizes, but no longer as 'raw' activity.

[3] the expected number of interactions for an individual is $d(\varepsilon_k)^{1-\beta_\varepsilon}$

Counterexample. (using (counterexample) and Equation 3.10)

$$\sum_k \mathbf{i}_{k1} = \sum_k \mathbf{i}_{k4}$$
$$\mathbf{w}(1,1) = \frac{1}{2^{2\beta_\varepsilon}} \neq \mathbf{w}(4,4) = \frac{2}{3^{2\beta_\varepsilon}}$$

↯

(*diagonal = max*) Satisfied

As mentioned, $\dfrac{\mathbf{i}_{ki}}{d(\varepsilon_k)}$ is constant for all i. Then since the set of elements that contribute to $\mathbf{w}(i,j)$ is the intersection of those that contribute to $\mathbf{w}(i,i)$ and $\mathbf{w}(j,j)$, the diagonal elements must necessarily be largest (but not strictly) in each column/row.

(*within-row/-column comparison*) Not Satisfied

By a similar argument as for **(*diagonal comparison*)**, interpreting comparisons for 'raw' activity is no longer valid.

Counterexample. (using (counterexample) and Equation 3.10)

$$\sum_k \mathbf{i}_{k1} \cdot \mathbf{i}_{k2} = \sum_k \mathbf{i}_{k2} \cdot \mathbf{i}_{k3}$$
$$\mathbf{w}(1,2) = \mathbf{w}(2,1) = \frac{1}{2^{2\beta_\varepsilon}} \neq \mathbf{w}(2,3) = \mathbf{w}(3,2) = \frac{1}{3^{2\beta_\varepsilon}}$$

↯

This edge weighting scheme already has most of the properties we want: it maintains **(*adding an interaction*)**, and breaks two out of the three 'bad' monotonicities of subsection 3.2.1. Normalising by interaction size proves to be a very effective way of creating a weighting function that contains more information than just co-observation, and that should not come as a surprise. This is where the extra information captured by $\mathbf{I}$ (as opposed to $\mathbf{A}$) becomes apparent.

3.3.3 'Mixed' Normalisation on I

In both previous normalisations, some monotonicity may have been lost due to the normalisation effect being duplicated in the multiplication for constructing $\mathbf{A}$. This is undesirable since it is an artefact of the operationalisation rather than a generic consequence of the

model. It is possible to remove this squaring effect by mixing row and column normalisations on $\mathbf{I}$. The resulting weights are then

$$\mathbf{w}(i,j) = \frac{\sum_k \frac{\mathbf{i}_{ki}}{d(\varepsilon_k)^{\beta_\varepsilon}} \cdot \mathbf{i}_{kj}}{d(v_j)^{\beta_{in}}} \tag{3.15}$$

Edge weights $\mathbf{w}(i,j)$ may still be interpreted as the proportion of interaction of v_j that is with v_i, as in subsection 3.3.1, but interaction is now measured as in subsection 3.3.2, giving less weight (expected interaction) to larger interactions (hyperedges).

Desirability The bounds on β_{in} and β_ε for this weighting scheme can be derived from earlier subsections: $\beta_{in}, \beta_\varepsilon < 1$.

(*adding an interaction*) NOT SATISFIED

Here, the generic counterexample cannot show what we would like it to, so a special counterexample is required.

Counterexample. If we start with one interaction of size 2 (ε_2), and one of size $n-1$ (ε_1), and add v_1 to ε_1

$$\mathbf{I} = \begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 0 \end{bmatrix}, \quad \mathbf{I}' = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 0 \end{bmatrix}$$

then $\mathbf{w}(2,1)$ should increase, but it does not:

$$\sum_k \mathbf{i}'_{k2} \cdot \mathbf{i}'_{k1} > \sum_k \mathbf{i}_{k2} \cdot \mathbf{i}_{k1}$$

$$\mathbf{w}(2,1) = \frac{1}{2^{\beta_\varepsilon}}, \quad \mathbf{w}(2,1)' = \frac{\frac{1}{2^{\beta_\varepsilon}} + \frac{1}{n^{\beta_\varepsilon}}}{2^{\beta_{in}}}$$

$$= 2^{-\beta_{in}} \cdot \left(\mathbf{a}_{21} + n^{-\beta_\varepsilon}\right)$$

$$\leq \mathbf{w}(2,1) \quad \text{if } \log_2\left(1 + \left(\frac{2}{n}\right)^{\beta_\varepsilon}\right) \leq \beta_{in}$$

and the strictness of both inequalities is linked. ↯

(*diagonal comparison*) NOT SATISFIED

As could be expected, this weighting scheme inherits the inability to compare diagonal elements as 'raw' activity from subsection 3.3.2 on page 42.

Counterexample. (using (counterexample) and Equation 3.10)

$$\mathbf{a}_{11} = \frac{1}{2^{\beta_\varepsilon}} \neq \mathbf{a}_{44} = \frac{1}{3^{\beta_\varepsilon}}$$

↯

(*diagonal = max*) Satisfied (column) Not Satisfied (row)

The diagonal element is the maximum of each column by a similar argument as in section 3.3.2 on page 43: $d\left(v_j\right)$ applies uniformly for all column elements, and the sum in the numerator is strictly monotonically increasing.

For the row maximum this scheme fails on the same counterexample as on the preceding page:

Counterexample.

$$\mathbf{I} = \begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ 1 & 1 & 0 & \dots & 0 \end{bmatrix}$$

$$\sum_k \mathbf{i}_{k2} = \sum_k \mathbf{i}_{k2} \cdot \mathbf{i}_{k1}$$

$$\mathbf{w}(2,2) = \frac{\frac{1}{2^{\beta_e}} + \frac{1}{(n-1)^{\beta_\varepsilon}}}{2^{\beta_{in}}} \leq \mathbf{w}(2,1) = \frac{1}{2^{\beta_\varepsilon}} \quad \text{if } \log_2\left(1 + \left(\frac{2}{n-1}\right)^{\beta_\varepsilon}\right) \leq \beta_{in}$$

↯

(*within-row/-column comparison*) Satisfied (column) Not Satisfied (row)

As mentioned on page 42, **(*diagonal = max*)** is a special case of **(*within-row/-column comparison*)**, and the same reasoning and counterexample apply.

A straightforward combination of row and column normalisation did not turn out well: although it did come close to being a complete departure from an element-wise mapping of $\mathbf{A}$, the price paid was in **(*adding an interaction*)**. Row and column normalisation need to be combined in a different way.

Alternative Operationalisation

The counterexample given for **(*adding an interaction*)** on page 45 relied on the change in the denominator of Equation 3.15 being disproportional to the change in the numerator. This suggests that the following may be a fruitful edge weighting to consider:

$$\mathbf{w}(i,j) = \frac{\sum_k \frac{\mathbf{i}_{ki} \cdot \mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}}{\left(\sum_\ell \frac{\mathbf{i}_{\ell j}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{in}}} \tag{3.16}$$

Now the numerator and denominator both count in the same units: a large interaction that contributes only weakly to the influence of v_i over v_j also only contributes weakly to v_j's 'quotum of influence', his/her total amount of influence received.

Desirability The properties of this weighting scheme are better than for the first proposal in this subsection because the column normalisation in the denominator is applied overtop of a row normalisation, harmonising currency between denominator and numerator. Again, monotonicity according to **(*adding an interaction*)** may become negative, but this can be resolved by requiring that $0 < \beta_\varepsilon(1 - \beta_{in}) < 1$.

(*adding an interaction*) Satisfied

Since we are only adding an interaction, say interaction k', between v_i and v_j (*ceteris paribus*), then the new weight $\mathbf{w}(i,j)'$ after adding the interaction will be

$$\mathbf{w}(i,j)' = \frac{\sum_k \frac{\mathbf{i}'_{ki} \cdot \mathbf{i}'_{kj}}{\left|\varepsilon'_k\right|^{\beta_\varepsilon}}}{\left(\sum_\ell \frac{\mathbf{i}'_{\ell j}}{\left|\varepsilon'_\ell\right|^{\beta_\varepsilon}}\right)^{\beta_{in}}} = \frac{\frac{1}{|\varepsilon_{k'}|^{\beta_\varepsilon}} + \sum_k \frac{\mathbf{i}_{ki} \cdot \mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}}{\left(\frac{1}{|\varepsilon_{k'}|^{\beta_\varepsilon}} + \sum_\ell \frac{\mathbf{i}_{\ell j}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{in}}}$$

and the denominator of Equation 3.16 always increases less rapidly than the numerator, and thus that $\mathbf{w}(i,j)$ is indeed monotonic with $\sum_k \mathbf{i}_{ki} \cdot \mathbf{i}_{kj}$, and strictly positively so if $\beta_{in} < 1$.

(*diagonal comparison*) Not Satisfied

the same counterexample as given on page 43 applies:

Counterexample. (using (counterexample) and Equation 3.10)

$$\mathbf{w}(1,1) = \left(\frac{1}{2^{\beta_\varepsilon}}\right)^{1-\beta_{in}} \neq \mathbf{w}(4,4) = \left(\frac{1}{3^{\beta_\varepsilon}}\right)^{1-\beta_{in}}$$

(when $\beta_{in} < 1$) ↯

(*diagonal = max*) SATISFIED

the diagonal element is the maximum of each column by a similar argument as on page 45: $\left[\sum_\ell \mathbf{i}_{\ell j} d(e_\ell)^{-\beta_\varepsilon}\right]^{\beta_{in}}$ applies uniformly for all column elements, and the sum in the numerator is strictly monotonically increasing. Perhaps surprisingly, the diagonal element is also the maximum of each row:

$$\mathbf{w}(i,i) = \left(\sum_k \frac{\mathbf{i}_{ki}}{d(\varepsilon_k)^{\beta_\varepsilon}}\right)^{1-\beta_{in}} \geq \left(\sum_k \frac{\mathbf{i}_{ki}\cdot\mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}\right)^{1-\beta_{in}} \geq \frac{\sum_k \frac{\mathbf{i}_{ki}\cdot\mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}}{\left(\sum_\ell \frac{\mathbf{i}_{\ell j}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{in}}}$$

(*within-row/-column comparison*) NOT SATISFIED

for within-column monotonicity, $\sum_\ell \mathbf{i}_{\ell j} d(\varepsilon_\ell)^{-\beta_\varepsilon}$ does not make any difference, so the counterexample from page 43 applies:

Counterexample. (using (counterexample) and Equation 3.10)

$$\mathbf{a}_{12} = \frac{\frac{1}{2^{\beta_\varepsilon}}}{\cancel{\left(\frac{1}{2^{\beta_\varepsilon}} + \frac{1}{3^{\beta_\varepsilon}}\right)^{\beta_{in}}}} \neq \mathbf{a}_{32} = \frac{\frac{1}{3^{\beta_\varepsilon}}}{\cancel{\left(\frac{1}{2^{\beta_\varepsilon}} + \frac{1}{3^{\beta_\varepsilon}}\right)^{\beta_{in}}}}$$

↯

and the same counterexample also works for row monotonicity.

The insight to "harmonise currency" was aimed primarily at monotonicity with respect to adding an interaction, but had the surprising side effect of also restoring the diagonal elements as maximum.

3.4 Joint Normalisation

Pulling all the strands from the previous schemes together, use the following to weight the edges:

$$\mathbf{w}(i,j) = \frac{\sum_k \frac{\mathbf{i}_{ki} \cdot \mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}}{\left(\sum_\ell \frac{\mathbf{i}_{\ell i}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{out}} \left(\sum_\ell \frac{\mathbf{i}_{\ell j}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{in}}} \tag{3.17}$$

Desirability Again, monotonicity according to **(*adding an interaction*)** may become negative, but this can be resolved by requiring that $0 < \beta_\varepsilon(1 - \beta_{out} - \beta_{in}) < 1$.

(*adding an interaction*) Satisfied

By a similar argument to the one on page 47, the denominator of Equation 3.17 always decreases less rapidly than the numerator, and thus $\mathbf{w}(i,j)$ monotonically increases when an interaction between v_i and v_j is added.

(*diagonal comparison*) Not Satisfied

the same counterexample as given on page 43 applies:

Counterexample. (using (counterexample) and Equation 3.10)

$$\mathbf{w}(1,1) = \left(\frac{1}{2^{\beta_\varepsilon}}\right)^{1-\beta_{in}-\beta_{out}} \neq \mathbf{w}(4,4) = \left(\frac{1}{3^{\beta_\varepsilon}}\right)^{1-\beta_{in}-\beta_{out}}$$

(when $\beta_{in} < 1$) ↯

(*diagonal = max*) Satisfied

Quite surprisingly, diagonal elements are still the maximum of each column and row:

$$\mathbf{w}(j,j) = \frac{\left(\sum_k \frac{\mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}\right)^{1-\beta_{out}}}{\cancel{\left(\sum_\ell \frac{\mathbf{i}_{\ell j}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{in}}}} \geq \frac{\left(\sum_k \frac{\mathbf{i}_{ki} \cdot \mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}\right)^{1-\beta_{out}}}{\cancel{\left(\sum_\ell \frac{\mathbf{i}_{\ell j}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{in}}}} \geq \frac{\sum_k \frac{\mathbf{i}_{ki} \cdot \mathbf{i}_{kj}}{d(\varepsilon_k)^{\beta_\varepsilon}}}{\left(\sum_\ell \frac{\mathbf{i}_{\ell i}}{d(\varepsilon_\ell)^{\beta_\varepsilon}}\right)^{\beta_{out}}}$$

(Swap β_{in} and β_{out}, and i and j to obtain the result for rows.)

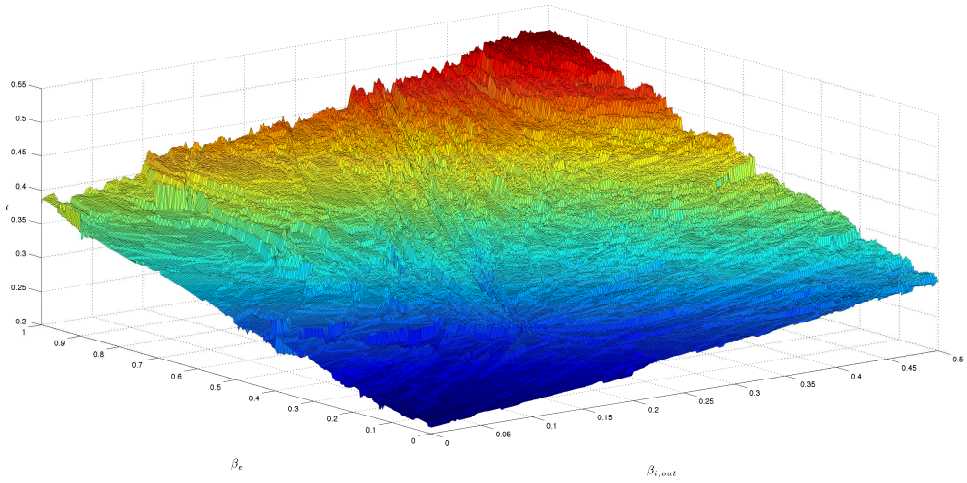


Figure 3.5: Parameter space of the edge weighting scheme of Equation 3.17

(***within-row/-column comparison***) NOT SATISFIED

The generic counterexample applies here as it did on page 47:

Counterexample. (using (counterexample) and Equation 3.10)

$$\mathbf{w}(1,2) = \frac{\frac{1}{2^{\beta_\varepsilon}}}{\left(\frac{1}{2^{\beta_\varepsilon}}\right)^{\beta_{out}} \cancel{\left(\frac{1}{2^{\beta_\varepsilon}} + \frac{1}{3^{\beta_\varepsilon}}\right)^{\beta_{in}}}} \neq \mathbf{w}(2,3) = \frac{\frac{1}{3^{\beta_\varepsilon}}}{\left(1 + \frac{1}{3^{\beta_\varepsilon}}\right)^{\beta_{out}} \cancel{\left(\frac{1}{2^{\beta_\varepsilon}} + \frac{1}{3^{\beta_\varepsilon}}\right)^{\beta_{in}}}}$$

↯

and the same counterexample also works for row monotonicity.

3.5 Application to ${}_{obs}\mathcal{H}^{prof}$

Proofs aside, our goal was and is to find an appropriate edge weight for ${}_{obs}G^{prof}$. Since modularity is undirected, we set $\beta_{in} = \beta_{out} = \beta_{in,out}$, and search the parameter space of Equation 3.17 for maximum modularity within the bounds $0 < 2\beta_{in,out}, \beta_\varepsilon \leq 1$. The surface representing modularity achieved by Newman's method [43] over this parameter space (shown in Figure 3.5) has many local maxima, but also a clear global maximum ($q = 0.515$) at ($\beta_{in,out} \approx 0.5$, $\beta_\varepsilon \approx 1$). The high modularity may be evidence that our edge weighting

scheme has uncovered some underlying social structure by adjusting for individuals' activity levels. However, it seems somewhat counterintuitive that β_ε should be close to 1. As explained in subsection 3.3.2, that would mean that every actor on average interacts with only a single person, no matter how large the meeting. From personal experience volunteering with the Cold Wet Weather Mat Programme, that seems unrealistic. In addition, actual optimal modularity is as high as $q = 0.863$, when $\beta_{in,out} \approx 1.5$, $\beta_\varepsilon \approx 4$, suggesting that the weighting scheme capitalises on odd cases, such as a single pair of actors who have only been seen once, together.

We conclude that the edge weighting scheme developed in this chapter, although well founded in principle, is not suited for application to these data. The edge weighting scheme has too strong a tendency to assign high importance to very low-frequency interactions. This could not be remediated by cutting low-frequency interactions, as each cut creates new low-frequency interactions (see Figure 3.2 and its explanation on page 33).

In the following, we will use the interaction frequency as edge weight $\mathbf{w}(i, j) = \mathbf{a}_{ij}$.

Chapter 4

Attribute-Based Edge Weights

The major assumption in applying the foregoing network analysis and arguing its usefulness is that outreach workers would have real-time access to the information derived from the analyses. While the Hope for Freedom Society's client group is small enough for the outreach workers to memorise up-to-date information (it may change over time) about each client in order to have it available when they are out on the street, that approach does not scale well.

To get around this obstacle, we attempt to essentialise within-community and between-community relationships using the clients' personal characteristics. In this way, information about relationships between particular individuals is translated into generic lessons about interactions between *types* of individuals whose personal characteristics are observable to the outreach workers.

The contention that it is indeed possible to essentialise some aspects of social interaction or relations between individuals to interactions or relations between generic actors with specific personal attributes is deeply entrenched in folk wisdom, for example in the sayings "birds of a feather flock together" and "opposites attract." It also lies at the basis of the hypothesis quoted in chapter 1, that communities form by drug preference. In the literature on social network analysis this idea has received attention in the form of network homophily (see reference [37] for an excellent review of evidence for homophily) and analyses using Exponential Random Graph Models (ERGMs; see Vol. 29, Iss. 2 of *Social Networks* for an overview).

Similarity In order to find out if similar actors interact more, 'similarity' needs to be defined. Since similarities for all attributes need to be meaningfully combined in the end,

define similarity over a particular attribute x between two actors v_i and v_j in a consistent and unitless way:

Categorical Attributes: (Drug of Choice and Sex)

$$\mathbf{sim}_x(i,j) = \begin{cases} 1 & \text{if } \mathbf{f}_x(v_i) = \mathbf{f}_x(v_j) \\ 0 & \text{otherwise (also if either is missing)} \end{cases} \tag{4.1}$$

where $\mathbf{f}_x(v_i)$ denotes personal feature (attribute) x of actor v_i. For example, if $\mathbf{f}_{sex}(v_i)$ = 'Male' and $\mathbf{f}_{sex}(v_j)$ = 'Female', then $\mathbf{sim}_{sex}(i,j) = 0$.

Categorised Continuous Attributes: (Age and Length of Homelessness)

$$\mathbf{sim}_x(i,j) = \begin{cases} 1 - \dfrac{\left|\mathbf{f}_x(v_i) - \mathbf{f}_x(v_j)\right|}{\max\limits_{i'} \mathbf{f}_x(v_{i'}) - \min\limits_{i'} \mathbf{f}_x(v_{i'})} & \text{normally} \\ 0 & \text{missing values} \end{cases} \tag{4.2}$$

where the category mean is used for $\mathbf{f}_x(v_i)$. For example, if $\mathbf{f}_{age}(v_i)$ = '< 33 years' and $\mathbf{f}_{age}(v_j)$ = '33–40 years', then $\mathbf{sim}_{age}(i,j) = 1 - |30\text{–}35|/50\text{–}30 = 3/4$.

Linear Regression As exploratory data analysis, we attempted a least squares linear regression. The model model we tried to fit would predict the number of times two actors have been observed together $\hat{\mathbf{a}}_{ij}$,[1] and contains a parameter β_x which goes with a linear term for similarity $\mathbf{sim}_x$ by each personal attribute x.

$$\begin{aligned} \hat{\mathbf{a}}_{ij} &= \beta_0 + \sum_x \beta_x \mathbf{sim}_x(i,j) \\ &= \beta_0 + \beta_{age}\mathbf{sim}_{age}(i,j) + \beta_{drug}\mathbf{sim}_{drug}(i,j) + \dots \end{aligned} \tag{4.3}$$

Least-squares regression minimises the sum of squared differences between the predicted ($\hat{\mathbf{a}}_{ij}$) and the observed ($\mathbf{a}_{ij}$) number of co-observations.

$$\min_{\beta_0, \beta_{age}, \dots} \sum_{i,j} \left(\mathbf{a}_{ij} - \hat{\mathbf{a}}_{ij}\right)^2$$

[1] if we had used presence or absence of a tie ($\mathbf{a}_{ij} \in \{0,1\}$), a logistic regression model would have been more appropriate.

It should be noted that this exploration was meant to gain preliminary insight into which personal attributes should be focused on. The model is not appropriate for the data because the data are not independently sampled.

MATLAB's STEPWISE-FIT function was used to conduct a stepwise regression. The algorithm proceeds as follows:

1. find the best (least squares) fit for an initial model containing only β_0.

2. estimate the p-value for each term that is currently not in the model.[2] Add the term with the lowest p-value less than 0.05, find the best fit, and repeat step 2.
   If no terms can be added, proceed to step 3.

3. remove the term with the largest p-value greater than 0.1, find the best fit, and go back to step 2.
   If no terms need to be removed, stop.

The goodness of fit of the model was evaluated using the coefficient of determination R^2:

$$R^2 = 1 - \frac{\sum_{i,j} \left(\mathbf{a}_{ij} - \hat{\mathbf{a}}_{ij}\right)^2}{\left(\mathbf{a}_{ij} - \bar{\mathbf{a}}\right)^2} \tag{4.4}$$

where $\bar{\mathbf{a}}$ is the mean over all $\mathbf{a}_{ij}$.

The best fitting model out of this procedure had too little explanatory power to be of any use ($R^2 = 0.005$). This may mean there is little association between personal attribute difference and co-observation, or that violations of the model assumptions were too severe. The fitting procedure assumes that all observed $\mathbf{a}_{ij}$ are sampled independently from an underlying distribution, *i.e.* that each sample is unrelated to any other sample. This is not the case: $\mathbf{a}_{ij}$ and $\mathbf{a}_{ik}$ are related because they share information about v_i. Consequently, the standard errors (and therefore p-values) are unreliable, undermining also the stepwise fitting procedure.

SIENA The precise effect of the violation of the independence assumption on the estimated β's and p-values depends on the structure of the network (the observed variables $\mathbf{a}_{ij}$), so we attempted to fit a model that takes into account network structure to correct

[2] the p-value of a term $\beta_x \left|\mathbf{f}_x(v_i) - \mathbf{f}_x(v_i)\right|$ is the probability that $\beta_x = 0$. A parameter β_x with a low p-value is *unlikely* to have *no* effect on the model.

for the violation of the independence assumption. We used SIENA (Simulation Investigation for Empirical Network Analysis) [51, 61], a programme that implements an estimation procedure for estimating an Exponential Random Graph Model (ERGM) [1, 39, 52, 60].

SIENA allows one to fit an equation much like Equation 4.3, except that extra terms for structural network effects (*e.g.* clustering) are added, and the error estimation is done by simulation. Because the error estimation is done by simulation, the fitting procedure may take quite long to reach any result (~ 1 week for our data). ERGMs are more fragile than regression models, and estimation may fail to converge. Indeed it did fail for even the simplest model, containing only a constant and a degree term. This means SIENA cannot be used to obtain a converging model [51], and we conclude that we are unable to find evidence of homophily at a person-to-person level.

Can the same be concluded at the community level? This question is precarious because its answer depends on how communities are defined and found.

4.1 Modularity

Motivated in part by our failure to find evidence for homophily at a person-to-person level, and in part by the low modularity afforded by ${}_{obs}G^{prof}$, we attempted to amend and strengthen the network community structure by adding information about similarity (**sim**) between actors' personal attributes to our original weights ($\mathbf{w}(i,j) = \mathbf{a}_{ij}$; as defined in section 3.5):

$$\mathbf{w}'(i,j) = \beta_{\mathbf{w}}\mathbf{w}(i,j) + \beta_{\mathbf{sim}}\mathbf{sim}(i,j), \qquad \beta_{\mathbf{w}}, \beta_{\mathbf{sim}} \geq 0 \tag{4.5}$$

In order to combine similarities across different personal attributes $\mathbf{f}_{age}$, $\mathbf{f}_{drug}$, *etc.*, use

$$\mathbf{sim}(i,j) = \sum_x \mathbf{sim}_x(i,j) \tag{4.6}$$

We do not use the β's estimated by the linear regression because they are rendered useless by the low R^2 value.

We would like to interpret $\beta_{\mathbf{w}}$ and $\beta_{\mathbf{sim}}$ as proportional contribution, so that $\beta_{\mathbf{w}} = 2\beta_{\mathbf{sim}}$ means that co-observation weights are indeed twice as important as similarity in features. This interpretation requires that $\mathbf{w}(\cdot,\cdot)$ and $\mathbf{sim}(\cdot,\cdot)$ use the same currency (operate at the same scale). To ensure this, normalise both **w** and **sim** by their standard deviation over all

possible vertex pairs:

$$\mathbf{w}'(i,j) = \beta_{\mathbf{w}} \frac{\mathbf{w}(i,j)}{\sigma_{\mathbf{w}}} + \beta_{\mathbf{sim}} \frac{\mathbf{sim}(i,j)}{\sigma_{\mathbf{sim}}}, \qquad \beta_{\mathbf{w}}, \beta_{\mathbf{sim}} \geq 0 \tag{4.7}$$

where $\sigma_{\mathbf{w}}$ is the standard deviation of the weights, and $\sigma_{\mathbf{sim}}$ is the standard deviation of the similarities. The standard deviation of a set of numbers X is defined as

$$\sigma(X) = \sqrt{\frac{1}{|X|-1} \sum_{x \in X} (x - \bar{x})^2}, \qquad \bar{x} = \frac{1}{|X|} \sum_{x \in X} x \tag{4.8}$$

Both $\mathbf{w}$ and $\mathbf{sim}$ will be symmetric functions, since modularity as we use it does not distinguish direction.

It is easy to see that a graph with weights according to Equation 4.7 will afford a community division with modularity at least 0.209, the maximum of the modularity for the original network ($q = 0.209$ with weights $\mathbf{w}(i,j)$) and the modularity for the homophily network ($q = 0.030$ with weights $\mathbf{sim}(i,j)$), achieved when $\beta_{\mathbf{w}} = 0$ or $\beta_{\mathbf{sim}} = 0$. The maximum modularity achievable between these two extremes ($\beta_{\mathbf{w}} = 0$ or $\beta_{\mathbf{sim}} = 0$) may provide some suggestion on whether homophily plays a role in community structure. As mentioned on the previous page, $\beta_{\mathbf{w}}$ and $\beta_{\mathbf{sim}}$ will be interpreted as proportional contribution. Also, the amount by which maximum modularity modularity increases from 0.209 will give an indication of the importance of homophily in community structure.

To find a vector of weights

$$\beta = \begin{bmatrix} \beta_{\mathbf{w}} \\ \beta_{\mathbf{sim}} \end{bmatrix}$$

that approximately maximise modularity, a community division ($\delta(i,j)$) needs to be given. Remember that Kronecker's $\delta(i,j)$ is 1 if v_i and v_j are in the same community, and 0 otherwise.

We cast the question of manipulating β to approximately maximise modularity as a quadratic programme with linear constraints. The general form for such a quadratic programme used by MATLAB's QUAD-PROG function is:

$$\min_{\beta} \beta^{\mathsf{T}} H \beta + h^{\mathsf{T}} \beta \text{ such that } \begin{cases} X\beta = y \\ lb \leq \beta \leq ub \end{cases} \tag{4.9}$$

In the following we derive H, h, X, y, lb and ub from B, the modularity matrix (see Equation 3.1).

The formulation of Equation 4.9 is quite removed from modularity as presented on page 30, since we now use the weight ($\mathbf{w}'(i,j)$) instead of 'raw' co-observation ($\mathbf{a}_{ij}$) to construct B:

$$b_{ij} = \left(\mathbf{w}'(i,j) - \frac{\left(\sum_{j'} \mathbf{w}'(i,j')\right)\left(\sum_{i'} \mathbf{w}'(i',j)\right)}{\sum_{i'}\sum_{j'} \mathbf{w}'(i',j')} \right) \tag{4.10}$$

So that modularity is now

$$q = \frac{\sum_{i,j} b_{ij}\delta(i,j)}{\sum_{i,j} \mathbf{w}'(i,j)} \tag{4.11}$$

To move toward the matrix notation of Equation 4.9, use

$$\mathbf{A}_{\mathbf{w}} = \begin{bmatrix} \frac{\mathbf{w}(1,1)}{\sigma_{\mathbf{w}}} & \cdots & \frac{\mathbf{w}(1,n)}{\sigma_{\mathbf{w}}} \\ \vdots & \ddots & \vdots \\ \frac{\mathbf{w}(n,1)}{\sigma_{\mathbf{w}}} & \cdots & \frac{\mathbf{w}(n,n)}{\sigma_{\mathbf{w}}} \end{bmatrix} \qquad \mathbf{A}_{\mathbf{sim}} = \begin{bmatrix} \frac{\mathbf{sim}(1,1)}{\sigma_{\mathbf{sim}}} & \cdots & \frac{\mathbf{sim}(1,n)}{\sigma_{\mathbf{sim}}} \\ \vdots & \ddots & \vdots \\ \frac{\mathbf{sim}(n,1)}{\sigma_{\mathbf{sim}}} & \cdots & \frac{\mathbf{sim}(n,n)}{\sigma_{\mathbf{sim}}} \end{bmatrix}$$

$$\mathbf{A}_{\mathbf{w}'} = \begin{bmatrix} \mathbf{w}'(1,1) & \cdots & \mathbf{w}'(1,n) \\ \vdots & \ddots & \vdots \\ \mathbf{w}'(n,1) & \cdots & \mathbf{w}'(n,n) \end{bmatrix} \qquad \mathbf{A}_{\delta} = \begin{bmatrix} \delta(1,1) & \cdots & \delta(1,n) \\ \vdots & \ddots & \vdots \\ \delta(n,1) & \cdots & \delta(n,n) \end{bmatrix}$$

. Now let the column vectors $D_{\mathbf{w}}$, $D_{\mathbf{sim}}$ and $D_{\mathbf{w}'}$ capture the degrees of the vertices:

$$D_{\mathbf{w}} = \mathbf{A}_{\mathbf{w}} e$$

and similarly for the others, where e is a column vector of ones of appropriate length. Now

$$e^{\mathsf{T}} D_{\mathbf{w}'} = \sum_{i,j} \mathbf{w}'(i,j)$$

so that modularity can be written as

$$q = \frac{1}{e^{\mathsf{T}} D_{\mathbf{w}'}} \left(\mathbf{A}_{\mathbf{w}'} - \frac{\beta_{\mathbf{w}}^2 D_{\mathbf{w}} D_{\mathbf{w}}^{\mathsf{T}} + \beta_{\mathbf{w}}\beta_{\mathbf{sim}} D_{\mathbf{w}} D_{\mathbf{sim}}^{\mathsf{T}} + \beta_{\mathbf{sim}}\beta_{\mathbf{w}} D_{\mathbf{sim}} D_{\mathbf{w}}^{\mathsf{T}} + \beta_{\mathbf{sim}}^2 D_{\mathbf{sim}} D_{\mathbf{sim}}^{\mathsf{T}}}{e^{\mathsf{T}} D_{\mathbf{w}'}} \right) \cdot \mathbf{A}_{\delta}$$

where $A \cdot B$ denotes the scalar product, or dot product, of A and B. It sums over the element-wise product of A and B.

Observe that $\mathbf{A}_\delta$ can be distributed

$$q = \frac{1}{e^\mathsf{T} D\beta}\left(\begin{bmatrix}\mathbf{A_w}\cdot\mathbf{A}_\delta & \mathbf{A_{sim}}\cdot\mathbf{A}_\delta\end{bmatrix}\beta - \frac{\beta^\mathsf{T}\begin{bmatrix}\left(D_\mathbf{w}D_\mathbf{w}^\mathsf{T}\right)\cdot\mathbf{A}_\delta & \left(D_\mathbf{w}D_\mathbf{sim}^\mathsf{T}\right)\cdot\mathbf{A}_\delta\\ \left(D_\mathbf{sim}D_\mathbf{w}^\mathsf{T}\right)\cdot\mathbf{A}_\delta & \left(D_\mathbf{sim}D_\mathbf{sim}^\mathsf{T}\right)\cdot\mathbf{A}_\delta\end{bmatrix}\beta}{e^\mathsf{T} D\beta}\right) \tag{4.12}$$

Now we have the start of a quadratic programme of the form of Equation 4.9 with

$$h = -\frac{1}{e^\mathsf{T} D\beta}\cdot\begin{bmatrix}\mathbf{A_w}\cdot\mathbf{A}_\delta & \mathbf{A_{sim}}\cdot\mathbf{A}_\delta\end{bmatrix},\quad H = \frac{1}{e^\mathsf{T} D\beta}\cdot\frac{\begin{bmatrix}\left(D_\mathbf{w}D_\mathbf{w}^\mathsf{T}\right)\cdot\mathbf{A}_\delta & \left(D_\mathbf{w}D_\mathbf{sim}^\mathsf{T}\right)\cdot\mathbf{A}_\delta\\ \left(D_\mathbf{sim}D_\mathbf{w}^\mathsf{T}\right)\cdot\mathbf{A}_\delta & \left(D_\mathbf{sim}D_\mathbf{sim}^\mathsf{T}\right)\cdot\mathbf{A}_\delta\end{bmatrix}}{e^\mathsf{T} D\beta}$$

From this arises a surprising insight: the size of the quadratic programme, and with it the computational cost of solving it, is independent of the number of vertices or edges, and quadratic in the number of attributes.

As constraints we can set $\beta_\mathbf{sim} = 1$, because only the relative weight of $\beta_\mathbf{w}$ with respect to $\beta_\mathbf{sim}$ is of interest to us, by the reasoning introduced on the previous page. Because all entries of B must be positive, one arbitrary but useful constraint would be that

$$e^\mathsf{T} D\beta = 1$$

to simplify H and h. This is permitted because scaling of all elements $\mathbf{a}_{ij}$ by the same amount does not affect modularity.

Proof. When applying a constant scaling factor c to $\mathbf{a}_{ij}$ so that $\mathbf{a}'_{ij} = c\cdot\mathbf{a}_{ij}$, b_{ij} changes as follows:

$$\begin{aligned}
b'_{ij} &= \frac{1}{2m}\left(\mathbf{a}'_{ij} - \frac{d'(v_i)\,d'(v_j)}{2m}\right)\\
&= \frac{1}{2\sum_{i,j}\mathbf{a}'_{ij}}\left(\mathbf{a}'_{ij} - \frac{\left(\sum_{j'}\mathbf{a}'_{ij'}\right)\left(\sum_{i'}\mathbf{a}'_{i'j}\right)}{2\sum_{i,j}\mathbf{a}'_{ij}}\right)\\
&= \frac{1}{2\cancel{c}\cdot\sum_{i,j}\mathbf{a}_{ij}}\left(\cancel{c}\cdot\mathbf{a}_{ij} - \frac{\cancel{c^2}\cdot\left(\sum_{j'}\mathbf{a}_{ij'}\right)\left(\sum_{i'}\mathbf{a}_{i'j}\right)}{2\cancel{c}\cdot\sum_{i,j}\mathbf{a}_{ij}}\right)\\
&= b_{ij}
\end{aligned}$$

□

The variables in Equation 4.9 now become:

$$H = \begin{bmatrix} \left(D_{\mathbf{w}} D_{\mathbf{w}}^{\mathsf{T}}\right) \cdot \mathbf{A}_{\delta} & \left(D_{\mathbf{w}} D_{\mathbf{sim}}^{\mathsf{T}}\right) \cdot \mathbf{A}_{\delta} \\ \left(D_{\mathbf{sim}} D_{\mathbf{w}}^{\mathsf{T}}\right) \cdot \mathbf{A}_{\delta} & \left(D_{\mathbf{sim}} D_{\mathbf{sim}}^{\mathsf{T}}\right) \cdot \mathbf{A}_{\delta} \end{bmatrix} \qquad h = -\begin{bmatrix} \mathbf{A}_{\mathbf{w}} \cdot \mathbf{A}_{\delta} & \mathbf{A}_{\mathbf{sim}} \cdot \mathbf{A}_{\delta} \end{bmatrix}$$

$$X = e^{\mathsf{T}} D \qquad y = 1 \qquad lb = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad ub = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

This quadratic programme finds a β that maximises modularity for a *given community division*. However, the community division is not given *a priori*; rather, it must be constructed in tandem with the optimisation. The method used so far does not lend itself well for such a task, because its step size is rather large: it proceeds by iteratively splitting communities in two until no more modularity gain can be achieved. To produce k communities this algorithm takes $k-1$ steps. An alternative, Newman's bottom-up approach [41] starts out with n singleton communities and proceeds by greedily merging the two communities that would result in the highest increase in modularity if they were merged, until no more modularity gain can be achieved. To produce k communities this algorithms takes $n-k$ steps. Usually in social networks, $k = \Theta(\log n)$ [41], giving the bottom-up approach a considerable edge over the top-down method in terms of resolution.

The quadratic programme of Equation 4.9 was 'nested' inside Newman's bottom-up algorithm [41]. After every merge of two communities, the quadratic programme was solved for the new community division using QuadProg. The new β, determined by QuadProg, was then used to determine B for the next step in Newman's bottom-up algorithm, *etc.* until modularity could no longer be increased. This combination (Newman's bottom-up algorithm with the quadratic programme nested inside it) was run 1000 times with random initial $\beta_{\mathbf{w}}$ ($\beta_{\mathbf{sim}}$ was set to 1).

4.2 Application to ${}_{obs}\mathcal{H}^{prof}$

The low modularity afforded by the 'raw' homophily network already foreshadowed that personal attributes might not have much to contribute to community structure. This is confirmed by an exploration of the parameter space of modularity-based community division using Equation 4.7 shown in Figure 4.1. The highest point ($q = 0.230$) lies at

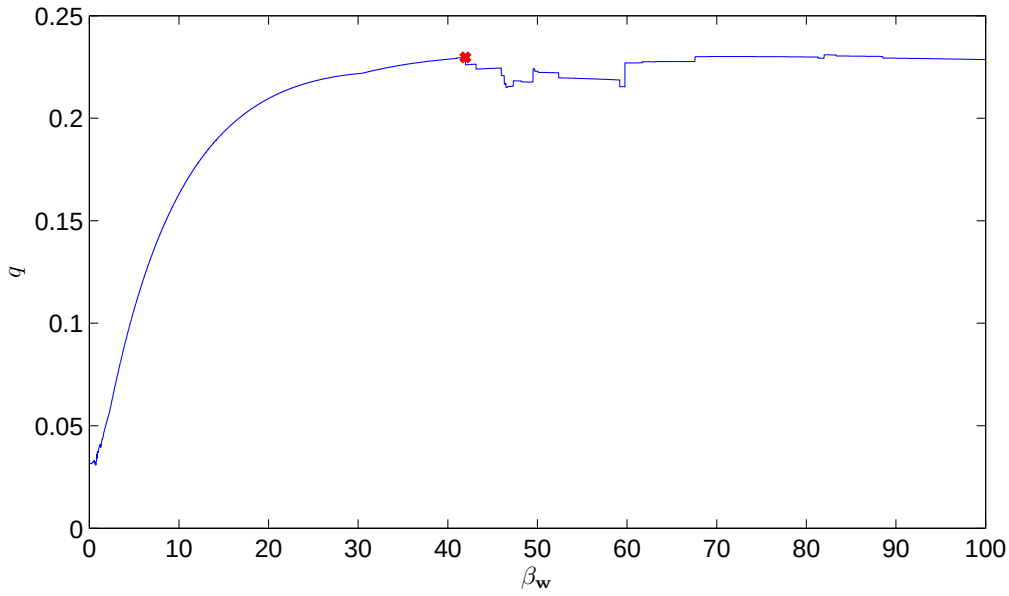


Figure 4.1: Modularity afforded by different relative weighting of co-observation and personal attributes in 4.5

$\beta_{\mathbf{w}} = 42.0$, $\beta_{\mathbf{sim}} = 1$, giving $\beta_{\mathbf{sim}}$ a rather small proportional contribution (2.33%) in $\mathbf{w}'$ for a comparatively large increase in modularity (9.90%).

The modularity contribution is not negligible, and we conclude that personal attributes may contribute to community structure, but that the contribution is small.

Chapter 5

Observation-Based Edge Weights

Two pieces of information that were extracted from the data have not yet been included in our edge weighting: locations and dates. As in the previous chapters, we incorporate these aspects to potentially strengthen the community structure as measured by modularity achieved by Newman's top-down method [43]. Location and time of observation do not necessarily add to the likelihood of interaction *per se*, but all the times and locations at which two individuals were co-observed do jointly influence the strength of the relationship between these two individuals. Therefore, both are taken as multiplicative factors L and T (defined on the following page and on the next page) that moderate relationship strength $\mathbf{w}'$ between vertices v_i and v_j rather than as additive terms:

$$\mathbf{w}^*(i,j) = L(i,j)T(i,j)\mathbf{w}'(ij) \tag{5.1}$$

where $\mathbf{w}'$ incorporates interaction and personal influence as defined in Equation 4.7.

Location The location at which interactions occurred may merit inclusion because particularly in the context of outreach, locations are not merely the environment within which interactions occur; they may be home to factors that actively attract or repel certain or all actors (*e.g.* a soup kitchen). This conceptualisation of the role locations play in interactions (and hence in observations) ties in well with the hint of location-based interaction found on page 31, in the fact that ${}_{obs}G^{loc}$ could be divided into 7 categories that were possibly representative of a distinction in the function of locations.

One could use the division into categories created by Newman's method to incorporate locations into edge weights by assigning a higher weight to a relationship between individuals

if they are co-observed at locations of diverse categories. It is possible, however, to extend this reasoning from categories to individual locations: intuitively, individuals that interact in few locations are less strongly connected than individuals that interact in many locations, *ceteris paribus*. Therefore, if $loc(v_i)$ denotes the set of locations at which v_i is observed, define the location factor L of co-observation as the Jaccard coefficient [30, 49] of the sets of locations at which v_i and v_j are observed:

$$L(i,j) = \frac{\left|loc(v_i) \cap loc\left(v_j\right)\right|}{\left|loc(v_i) \cup loc\left(v_j\right)\right|} \tag{5.2}$$

A high location factor ($L(i,j) = 1$) is achieved by a pair of individuals whose activity spaces (sets of location they visit) are identical, and a low location factor ($L(i,j) = 0$) is achieved by a pair of individuals whose activity spaces do not overlap at all.

Time The time at which interactions occurred may merit inclusion because it provides information on whether or not the relationship between individuals is a sustained one or not. Individuals who have interacted over a longer period of time likely have a stronger relationship than individuals who have interacted over a shorter period of time.

If $t(v_i)$ denotes the set of dates at which v_i is observed, then define the time factor of co-observation akin to Equation 5.2, but using the standard deviation (σ; see Equation 4.8) as a measure of spread:

$$T(i,j) = \frac{\sigma\left(t(v_i) \cap t\left(v_j\right)\right)}{\sigma\left(t(v_i) \cup t\left(v_j\right)\right)} \tag{5.3}$$

The reasoning for not using the Jaccard coefficient (Equation 5.2) directly is that the dates are continuous, and hence provide more information than simple set overlap. A high time factor is achieved by a pair of individuals who are observed together consistently over time. A low time factor is achieved by a pair of individuals who interact in only a short period of time, but are separately observed for a much longer period of time.

This comparison assumes that all individuals can interact with one another from the very first day of recording. If a large group of people only entered the population at a later date, the duration of a relationship would be influenced this later entry. However, although some clients were indeed not yet homeless when the Hope for Freedom Society first started their outreach activities, the majority was (> 80% by January '08, with only

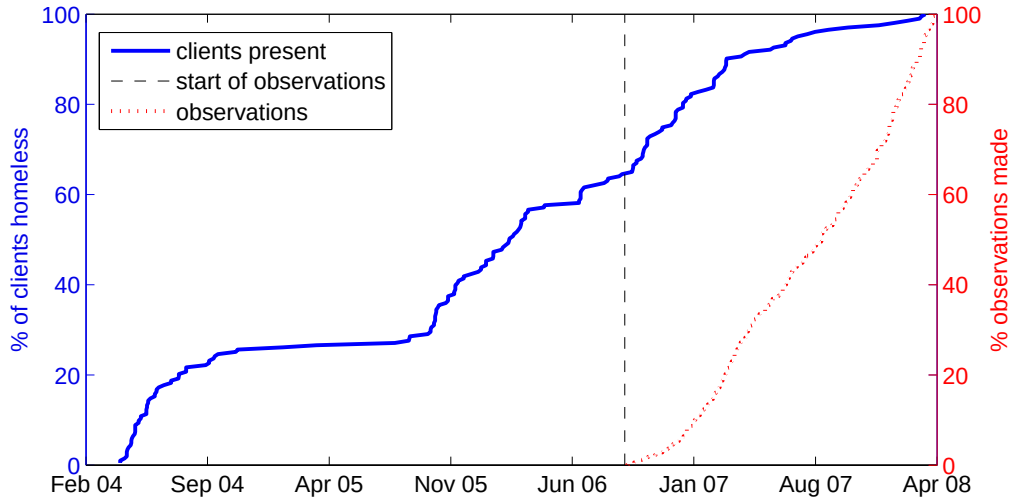


Figure 5.1: Percentage of clients that is homelessness (53 values missing) and percentage of observations made over time (cumulative). The vertical dashed line represents when the Hope for Freedom Society started their outreach activities

< 10% of observations made; see Figure 5.1),[1] so this should be only a minor concern.

5.1 Application to ${}_{obs}\mathcal{H}^{prof}$

These two factors have the potential to affect modularity to quite a large extent, because individually they already afford a comparatively high modularity ($q = 0.102$ for the location factor, and $q = 0.272$ for the time factor). In fact, it seems that the time factor dominates $\mathbf{w}^*$, because the resulting modularity is almost exactly the same ($q = 0.272$). The community divisions, however, are quite different: the community division by the time factor (δ_T) results in 7 communities, whereas community division by $\mathbf{w}^*$ ($\delta_{\mathbf{w}^*}$) results in 15 communities.

The adjusted Rand index [28], which compares two partitions of the same set, is used to compute how similar the community divisions are. If c_{ij} is the number of vertices of community i in δ_T that are also in community j of $\delta_{\mathbf{w}^*}$, and $c_{i\cdot}$ and $c_{\cdot j}$ the total number of vertices in community i in δ_T and community j of $\delta_{\mathbf{w}^*}$, respectively, then the adjusted

[1] we assume that once an individual becomes homeless they are able to interact with other clients. The date of commencement of homelessness is calculated as the earliest date at which a person was observed minus the length of homelessness recorded in his/her file.

Rand index is defined as

$$R(\delta_T, \delta_{\mathbf{w}^*}) = \frac{\sum_{i,j}\binom{c_{ij}}{2} - E\left(\binom{c_{ij}}{2}\right)}{\left(\text{upper bound on } \sum_{i,j}\binom{c_{ij}}{2}\right) - E\left(\binom{c_{ij}}{2}\right)}$$
$$= \frac{\sum_{i,j}\binom{c_{ij}}{2} - \frac{\sum_i \binom{c_{i\cdot}}{2}\sum_j \binom{c_{\cdot j}}{2}}{\binom{n}{2}}}{\frac{\sum_i \binom{c_{i\cdot}}{2}\sum_j \binom{c_{\cdot j}}{2}}{2} - \frac{\sum_i \binom{c_{i\cdot}}{2}\sum_j \binom{c_{\cdot j}}{2}}{\binom{n}{2}}}$$

and ranges from 0 to 1. For the community divisions by T and $\mathbf{w}^*$, the adjusted Rand index is low: $R(\delta_T, \delta_{\mathbf{w}^*}) = 0.092$, confirming that the community divisions are different, and thus that the modularity was not a result of the time factor dominating the behaviour of $\mathbf{w}^*$.

5.2 Is Something Social Going On?

The results in the foregoing have not been what should be expected from a social network. Therefore we ask "is something social going on?" Nothing so far has suggested a positive answer. In this section, we make a final attempt at discovering something social, investigating if there are at least social *locations*, which we define as those locations people meet at for social reasons. The motivation behind this is perhaps most easily understood through a negative example: a soup kitchen is not expected to be a social location (although much social interaction may occur over a hearty meal), since a client's motivation for being at this location is largely the *service* provided there rather than the *people* who are there. In cases where the location is not the primary reason for a client being at that location, co-observation of multiple clients may indicate that something social is going on.

5.2.1 Methods

In order to test if co-observations of two individuals at a certain location are evidence of a social connection between them, we compute the probability that nothing social is going on. For this we use a null model in which clients go to locations and meet unintendedly in the process.

Let X be a random variable denoting the number of meetings between person 1 and person 2, then the probability distribution of X

$$\Pr(X \mid n, n_1, n_2), \tag{5.4}$$

depends on the total number of observations (n) at this location, the number of observations of person 1 (n_1) and the number of observations of person 2 (n_2). Naturally, $n_1, n_2 \leq n$.

Then if person 1 and 2 are seen together k times at this location, we wish to compute the probability that this is positively unusual

$$\Pr(X \geq k \mid n, n_1, n_2) = \sum_{i=k}^{\min(n_1,n_2)} \Pr(X = i \mid n, n_1, n_2) \tag{5.5}$$

This probability can be rewritten discretely as

$$\Pr(X = k \mid n, n_1, n_2) = \frac{N(n, n_1, n_2, k)}{\binom{n}{n_1}\binom{n}{n_2}}, \tag{5.6}$$

where $N(n, n_1, n_2, k)$ is the number of possible meeting patterns in which person 1 and person 2 are observed together exactly k times. It may be defined recursively as

$$\begin{aligned} N(n, n_1, n_2, k) = \quad & N(n-1, n_1, n_2, k) && \text{[neither observed]} \\ & + N(n-1, n_1-1, n_2, k) && \text{[person 1 observed]} \\ & + N(n-1, n_1, n_2-1, k) && \text{[person 2 observed]} \\ & + N(n-1, n_1-1, n_2-1, k-1) && \text{[both observed]} \end{aligned}$$

The ubiquitous "$n-1$" term points to redundancy in this recursive formulation, which can be removed. Define

$$\begin{aligned} n_{11} &= k \\ n_{10} &= n_2 - k \\ n_{01} &= n_1 - k \\ n_{00} &= n - (n_{00} + n_{01} + n_{10}) \\ &= n - n_1 - n_2 + k \end{aligned}$$

so that

$$\begin{aligned} N'(n_{00}, n_{01}, n_{10}, n_{11}) = \quad & N(n_{00} + n_{01} + n_{10} + n_{11}, n_{01} + n_{11}, n_{10} + n_{11}, n_{11}) \\ = \quad & N'(n_{00} - 1, n_{01}, n_{10}, n_{11}) && \text{[neither observed]} \\ & + N'(n_{00}, n_{01} - 1, n_{10}, n_{11}) && \text{[person 1 observed]} \\ & + N'(n_{00}, n_{01}, n_{10} - 1, n_{11}) && \text{[person 2 observed]} \\ & + N'(n_{00}, n_{01}, n_{10}, n_{11} - 1) && \text{[both observed]} \\ = \quad & \binom{n_{00} + n_{01} + n_{10} + n_{11}}{n_{00}, n_{01}, n_{10}, n_{11}} \end{aligned}$$

The multinomial has a clear interpretation: it is the number of distinct sets of n $\{0,1\}$-strings (events) that contain exactly k co-observations, $n_1 - k$ observations of person 1 alone, $n_2 - k$ observations of person 2 alone, and the remaining observations of neither. Given the large number of probability queries, the recurrence will be used in dynamic programming to compute the probabilities instead of the multinomial.

Now Equation 5.6 can be written as

$$\Pr(X = k \mid n, n_1, n_2) = \frac{\binom{n}{n-n_1-n_2+k, n_1-k, n_2-k, k}}{\binom{n}{n_1}\binom{n}{n_2}}, \tag{5.7}$$

This rewriting also makes it possible to rewrite Equation 5.5, since the denominator is independent of k:

$$\Pr(X \geq k \mid n, n_1, n_2) = \frac{\sum_{i=k}^{\min(n_1,n_2)} N'(n - n_1 - n_2 + i, n_1 - i, n_2 - i, i)}{\binom{n}{n_1}\binom{n}{n_2}} \tag{5.8}$$

Therefore, define:

$$\begin{aligned} N''(n_{00}, n_{01}, n_{10}, n_{11}) &= \sum_{i=0}^{\min(n_{01},n_{10})} N'(n_{00} + i, n_{01} - i, n_{10} - i, n_{11} + i) \\ &= N'(n_{00}, n_{01}, n_{10}, n_{11}) + N''(n_{00} + 1, n_{01} - 1, n_{10} - 1, n_{11} + 1) \end{aligned}$$

with boundary condition

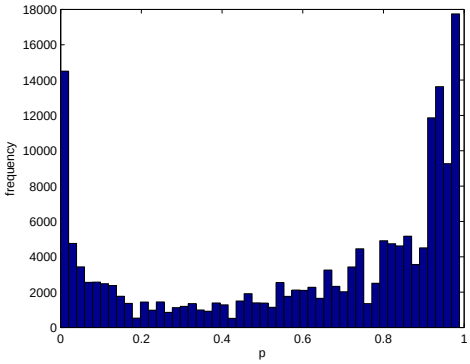
$$\begin{aligned} \forall n_{00}, n_{10}, n_{11} \;\; N''(n_{00}, 0, n_{10}, n_{11}) &= N'(n_{00}, 0, n_{10}, n_{11}) \\ \forall n_{00}, n_{01}, n_{11} \;\; N''(n_{00}, n_{01}, 0, n_{11}) &= N'(n_{00}, n_{01}, 0, n_{11}) \end{aligned}$$

to compute $\Pr(X \geq k \mid n, n_1, n_2)$ efficiently using dynamic programming.

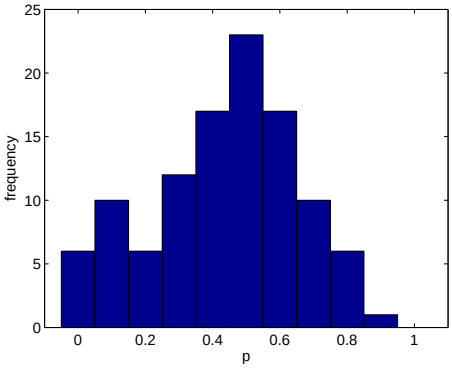
5.2.2 Results

If the null model were appropriate and true, the p-values of Equation 5.5 should be uniformly distributed. A higher probability mass at lower values would indicate that something social is going on. The distribution of p-values shown in Figure 5.2 clearly suggests that indeed there is no evidence of *positive* social interaction under the defined null model. No locations seem to be especially social or anti-social (see Subfigure 5.2(b)), since the average p-values for

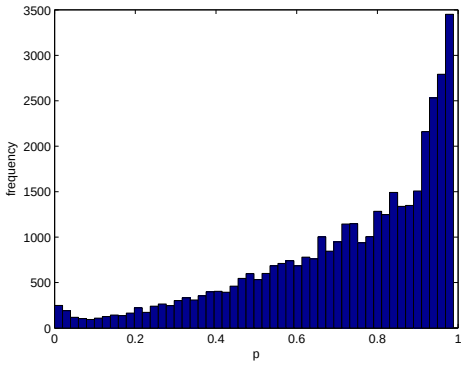
locations appear to be almost normally distributed, but they are not ($p \ll 10^{-10}$ Kolgomorov-Smirnov). The appearance is due to the binning done by the histogram. We used fewer bins (bars) for this plot due to the low number of samples ($N = 150$) in comparison with the others ($N = 256^2$ and $N = 256^2 \times 150$). The skew of average actor-to-actor p-values (in figure Subfigure 5.2(c)) towards higher p-values may even be indicative of negative social interaction (avoidance behaviour), but it is more likely an indication that the null model is inappropriate. If, for example, either or both individuals were not able to freely (randomly) appear at a given location, the expected number of interactions should be lower than it is under this null model, and consequently the p-value would be overestimated. Evidence of this is visible in both Subfigure 5.2(a) and Subfigure 5.2(c). The spike at 0 in figure Subfigure 5.2(a) is due to observations for which $n_1 = n_2 = k$, where usually $k = 1$ or at least small. The fact that there is a spike means that such instances occur more than would be expected. These instances are, however, not centered on any pair of individual or location, so we cannot say they are due to a relationship or a location conducive to social interaction.



(a) for each (observation, actor, actor) triple



(b) average per location



(c) average per actor pair

Figure 5.2: p-values for Equation 5.5: the probability that interaction between two individuals at a location is random

Chapter 6

Conclusion

> Can we find evidence of something social going on among homeless in the Tri Cities?

In search of an answer to this question, we used data provided by the Hope for Freedom Society, a homeless outreach organisation. These data, raw text logs of interaction between outreach workers and homeless clients, were anonymised (section 2.1), and information about social interaction among homeless in the Tri Cities was extracted from them (section 2.2). The resulting social networks were then analysed for evidence of community structure (section 3). A principled way of determining the strength of the relationship between two individuals was developed, based on interaction (chapter 3), personal characteristics (chapter 4) and personal activity spaces (chapter 5), and at each step the resulting weighted network was analysed for evidence of community structure.

Despite the inclusion of interaction information, personal information, locations and dates, we were unable to find evidence of community structure or positive social interaction in the data. The interaction information shows signs of a core-periphery structure, suggesting that there may be just a single community. The hypothesis mentioned in chapter 1, that the community is subdivided by drug preference, should therefore likely be rejected, based on these data. It may be argued, as in Figure 2.1, that the data on drug preference does not reflect actual drug use, and that this lies behind our inability to find drug-based communities. This is, however, to some extent also the case in the context from which the hypothesis arises. Moreover, the Hope for Freedom Society is first and foremost a drug rehabilitation organisation, so their record of a person's drug preference is likely as accurate

as possible.

The addition of personal attributes to the information captured in edge weights allowed modularity to increase compared to using just co-observation frequency as edge weight. Although the increase was quite substantial (8.61%), the resulting modularity is still well below the range expected for a social network ($0.4 \leq q \leq 0.7$) [7, 12, 43]. The inclusion of time and location overlap between pairs of actor (in chapter 5) did increase modularity (to $q = 0.272$), but it was still well outside of the 'social' range. It did become clear that time played an important role in the network, and it would be interesting to see if time-evolving communities could be found. Some modularity-based methods for time-evolving communities have been developed [40], but at present these only use time *slices* rather than continuous time.

However, if nothing by way of social interaction is present in the data, little can be extracted. We have been unable to find evidence of positive social interaction in the data, but our most direct approach to this (in section 5.2) relies on a null model in which individuals can freely partake in any interaction at locations they have visited at least once. Similarly, the null model underlying modularity assumes that each individual could interact freely with each other individual. Geographical constraints and absences (*e.g.* hospitalisation, drug rehabilitation, housing) could undermine these assumptions. The way observations are made also has its inherent limitations: clients do not freely participate in an observation made as a result of an emergency call, for example. Thus our inability to find evidence of positive social interaction may be due to limitations of our methods.

Taking the (absence of) evidence for what it is, the fact that we were unable to find evidence of positive social interaction may be a positive sign. If social relationships among homeless are a significant factor in maintaining homelessness [47], then the Hope for Freedom Society is positively contributing to resolving homelessness by not providing venues for social interaction. The focus of their outreach is on the service provided [66], and social interaction is mostly geared towards that. The factors that alleviate homelessness are already present in the Hope for Freedom Society's outreach, so the need for social interaction in this respect may be less than reported in reference [67].

Much more information is present in the data than could not be extracted, particularly in the semantics of the interaction logs, but also just in the accuracy of the data. The Hope for Freedom Society could increase the effectiveness of their data collection, and thereby the

usefulness of their data by analysing it *in house*.[1] Although our anonymisation approach could be applied as data are collected, it would be even better if the tagging of individuals in an observation used the outreach workers' input. One could, for example, use auto-completion to link names of individuals and locations to personal profiles and geocoded positions. Such a feature could be seemlessly embedded in the interface used for logging activities, so that the increase in workload for the outreach workers is kept to a minimum. With such small adjustments, the data could become useful in *real time* (not after 2 years of a Master's thesis). For example, the data could be used to detect disappearances, or for internal or external audit.

[1]The data we used is exceptionally detailed, and a good example of record keeping among other outreach and government organisations. The Hope for Freedom Society's commitment to data collection is commendable, especially with the small number of staff it has. The suggestions and recommendations made should be taken as encouragement rather than criticism.

Appendix A

Anonymisation Procedure

A.1 Introduction

We are proposing a methodology for systematically substituting identifiers (*e.g.* `<<N5468>>`) for names in textual data sets, without anyone looking at any semantically meaningful part of the data. Our proposal is aimed at removing 99.9% of names while minimising semantic information loss.

A.2 Procedure

For this procedure we would employ three people who would not be informed about the source of the data they are working with.

1. mechanically process the data
    - extract all unique words of alphabetical characters from the data. numbers, apostrophes, punctuation and any other semantics beyond the word level are ignored, but capitalisation is preserved.
    - write these unique words into a two-column MS Excel file in alphabetical order, and fill the second column with `Ignore`.
2. Identify the names and possible names
   three people perform this task independently of one another, and the resulting lists are compared

- words that are uniformly identified as names are added to the dictionary of names
- other words that are flagged at least once as a word that cannot be ignored are added to the dictionary of possible names

Both are written to a MS Excel file for the researcher (or a fourth employee) to disambiguate.

3. Group similar names
   three people perform this task independently of one another, and the resulting lists are compared by the researcher (or a fourth employee).

The alphabetical order destroys any correlation between words, and with it all semantics beyond the word level.

A.3 Full MATLAB code

Listing A.1: Extract Words

```
function [words] = extractWords( fileName )
    % open the file for reading
    fid = fopen( fileName, 'r' );

        % measure the file size
        D = dir( fileName );
        fileSize = D.bytes;

        % read the file into a sufficiently large string
        text = textscan( fid, '%s', 'Delimiter', '',...
                                    'BufSize', fileSize+1 );
        % split the text at non-alphabetical characters.
        words = regexp( text, '[\W^\d]+', 'split' );

    fclose( fid );
end
```

Listing A.2: Write Excel "Dictionary"

```
function writeDictionary( fileName, words )
    % open the file for writing
    fid = fopen( fileName, 'w' );

        % write the words to a tab-delimited file
        fprintf( fid, '%s\tIgnore\r\n', words{:} );

    fclose( fid );
end
```

Listing A.3: Parse Excel "Dictionary"

```
function [dicts] = parseDictionary( inputFileName )
%% READ FROM THE FILE
    % open the file for reading
    fid = fopen( inputFileName, 'r' );

        % the file must again be tab-delimited
        columns = textscan( fid, '%s %s', 'Delimiter', '\t',...
                                          'endOfLine', '\r\n' );
        % the first column contains the words
        words = columns{1};
        % the second column contains the categorisations;
        %   ignore lowercase/uppercase differences.
        cats = lower( columns{2} );

    fclose( fid );

%% INITIALISE THE OUTPUT
    dicts = struct;
    % create (currently empty) dictionaries per category
    for category = unique(cats)'
        % initialise empty dictionary
        dicts.(category{:}) = cell(0);
    end

%% PARSE THE FILE
    % for all words
```

```
    for w = 1:numel(words)
        % append this word to the appropriate dictionary
        dicts.( cats{x} ){end+1} = words{x};
    end
end
```

Listing A.4: Compare Excel "Dictionaries"

```
function compareDictionaries( dictFileNames, outputFileName )
%% READ FROM THE FILE
    % initialise data storage
    words = cell( 0 );
    cats = cell( 1, numel(dictFileNames) );
    for f = 1:numel(dictFileNames)
        % open the file for reading
        fid = fopen( dictFileNames{f}, 'r' );

            % the file must again be tab-delimited
            columns = textscan( fid, '%s %s', 'Delimiter', '\t',...
                                              'endOfLine', '\r\n' );
            % the first column contains the words
            words = columns{1};
            % the second column contains the categorisations;
            %   ignore lowercase/uppercase differences.
            cats{f} = lower( columns{2} );

        fclose( fid );
    end

    % concatenate the categorisations
    cats = [ cats{:} ];

    % sort by categorisation
    [cats,indices] = sortrows( cats, 1:size(cats,2) );
    % also permute the words accordingly
    words = words(indices);

%% COMPARE DICTIONARIES
    % open the file for writing
```

```
32      fid = fopen( outputFileName, 'w' );
33
34          numberOfColumns = size( cats, 2 );
35          lineForAmbiguousInput = [ '%s',...
36                          repmat('\t%s',1,numberOfColumns) '\r\n' ];
37
38          % write the words to a tab-delimited file
39          for w = 1:numel(words)
40              if( all( ismember(cats(w,:),cats{w,1}) ) )
41                  % only need to write a single category to the file
42                  fprintf( fid, '%s\t%s\r\n', words{w}, cats{w,1} );
43              else
44                  fprintf( fid, lineForAmbiguousInput,...
45                               words{w}, cats{w,:} );
46              end
47          end
48
49      fclose( fid );
50  end
```

Listing A.5: Write Excel "Thesaurus"

```
1   function writeThesaurus( fileName, thesaurus )
2       % open the file for writing
3       fid = fopen( fileName, 'w' );
4
5           % write the words to a tab-delimited file
6           fprintf( fid, '%s\t%s\r\n', thesaurus{1,:},...
7                                       thesaurus{2,:} );
8
9       fclose( fid );
10  end
```

Listing A.6: Parse Excel "Thesaurus"

```
1   function [thesaurus] = parseThesaurus( inputFileName )
2       % open the file for reading
3       fid = fopen( inputFileName, 'r' );
```

```
4
5           % the file must again be tab-delimited
6           columns = textscan( fid, '%s %s', 'Delimiter', '\t',...
7                                             'endOfLine', '\r\n' );
8           % the first column contains the words
9           words = columns{1};
10          % the second column contains the categorisations;
11          %   ignore lowercase/uppercase differences.
12          substitutions = lower( columns{2} );
13
14      fclose( fid );
15
16      thesaurus = sortrows( [ substitutions, words ] );
17  end
```

Listing A.7: Anonymise

```
1   function anonymise( fileName, dicts, thesaurus )
2       % open the file for reading and writing
3       fid = fopen( fileName, 'r+' );
4
5           % measure the file size
6           D = dir( fileName );
7           fileSize = D.bytes;
8
9           % read the file into a sufficiently large string
10          text = textscan( fid, '%s', 'Delimiter', '',...
11                                      'BufSize', fileSize+1 );
12
13          % do thesaurus substitutions first
14          text = strrep( text, thesaurus{1,:}, thesaurus{2,:} );
15          % do replacement for every dictionary
16          for dict = fieldnames(dicts)'
17              % substitute numbered codes for the words in
18              %   this dictionary
19              subst = ( 1:numel(dicts.(dict{:})) )';
20              subst = [ '<<' dict{:} num2str(subst) '>>' ];
21              text = strrep( text, dicts.(dict{:}){:}, subst );
22          end
```

```
23
24          % write the anonymised text back to the file
25          fprintf( fid, '%s', text );
26
27      fclose( fid );
28  end
```

Appendix B

Conceptual Model

> "[A]ll any model is supposed to do [...] is to provide an abstract representation of effects that are important in determining the behavior of a system. And below the level of these effects there is no reason that the model should actually operate like the system itself." [70, p. 366]

In this chapter we will introduce graphs in a quite general form and explore how they may be appropriate to model particular social phenomena. For a less broad but more detailed introduction to Social Network Analysis, please refer to the books used as resources for this section [34, 24, Ch. 3]. Good introduction to graphs can be found in [8, Ch. 1], [29, Ch. 2].

B.1 Graphs

Graphs are a collection of *objects*, *actors* or *entities*, represented and referred to as *vertices*, which are associated with each other, typically through *interactions* or *relations*, represented and referred to as *edges*; graphs represent things that somehow have something to do with other things.

As defined here, graphs represent only the descriptive *topology* (structure) of relationships between the objects represented as vertices; there is no mention of information about the objects or relations. We will ignore such additional information as we guide the reader through a series of graphs with ever more general definitions of an edge, introducing the requisite terminology as we move along. First we consider the number vertices participating in an edge and then the possibility of duplication.

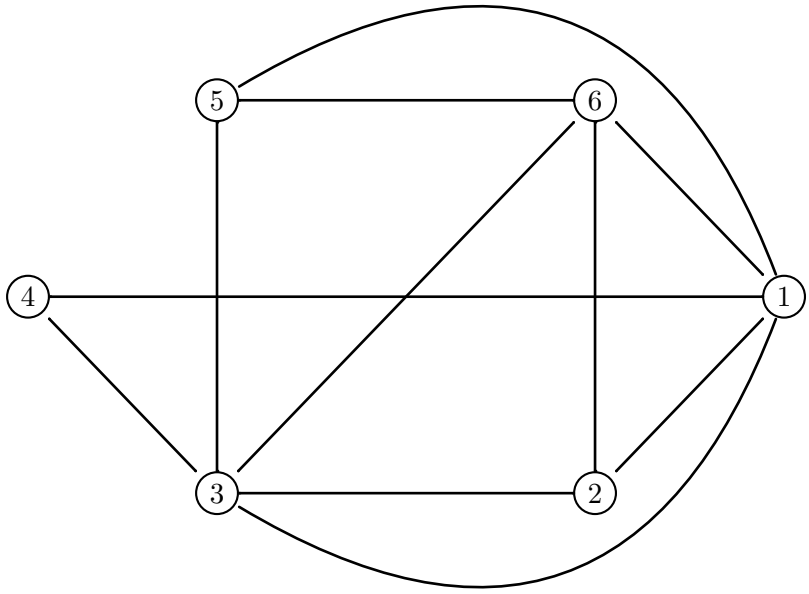


Figure B.1: An example simple graph

B.1.1 Simple Graphs

Our base case for this exposition is the *simple graph*. In a simple graph, edges represent symmetric bilateral association: they are (unordered) pairs of vertices. In this case, where edges always relate exactly two vertices, it is common to use e_{ij} (or e_{ji}, by symmetry of the relationship) to denote an edge relating v_i and v_j. A good example of a simple graph would have facebook© profiles as vertices and approved friend links as edges (see Figure B.1 for a toy example).

More precisely, $E \subseteq \wp_2(V)$, where the notation $\wp_k(V)$ denotes the set (unordered collection) of all sets of k elements drawn from a set V:

$$\wp_k(V) = \left\{ e \subseteq V \;\middle|\; |e| = k \right\}$$

This definition of edges as sets immediately implies the restriction that there be no *loops*, *i.e.* that a vertex cannot be in a relation with itself: $e_{ii} = \text{`}\{v_i, v_i\}\text{'} = \{v_i\} \notin \wp_2(V)$. This restriction on loops makes sense in the example of facebook© friendship links, but not in a graph that has people as vertices and kinship links upto the second degree as edges: you are your own kin.

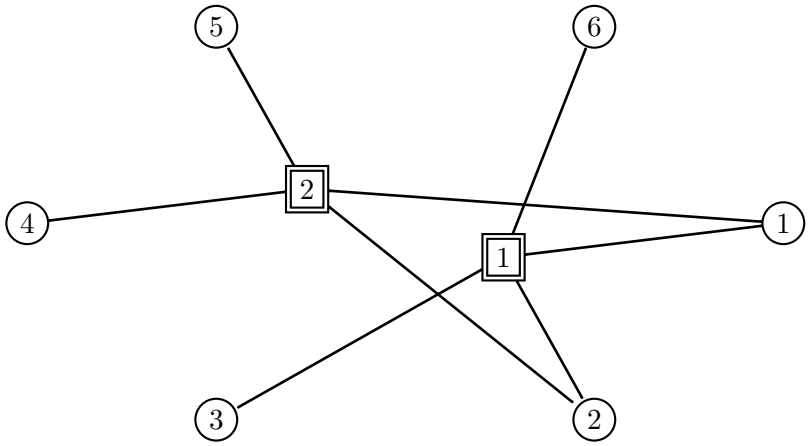


Figure B.2: An example hypergraph (circles are vertices, boxes are hyperedges)

B.1.2 Hypergraphs

Facebook© friendship is a necessarily bilateral association, but for example bobsleigh 4-crew membership is necessarily quadrilateral (see Figure B.2), and institutional affiliation is multilateral or oligolateral: relations, and therefore edges, may have more than two participants, and the number of participants may vary between relations. Graphs with such edges are called *hypergraphs*, usually denoted $\mathcal{H}$.

A hyperedge $\varepsilon \in \mathcal{E}$ is a set of vertices, so $\mathcal{E} \subseteq \wp(V)$ without limit on the size of the edges.

B.1.3 Duplication

When edges model interactions rather than relationships, the same edges may typically occur multiple times. The above definition does not allow this because E and $\mathcal{E}$ are sets, enforcing the uniqueness of each edge. Defining E or $\mathcal{E}$ as a multiset (unordered collection allowing duplicates) lifts this restriction: $E, \mathcal{E} \subseteq \big(\wp(V) \times \mathbb{N}\big)$, so an edge is a pair $\big(e, \mathbf{1}_E(e)\big)$, $e \subseteq V$ or $\big(\varepsilon, \mathbf{1}_\mathcal{E}(\varepsilon)\big)$, $\varepsilon \subseteq V$ of the 'actual' edge (which is a set of vertices), and a number indicating the multiplicity of the edge, *i.e.* how many copies of this edge exist in the graph. $\mathbf{1}_E$ is the usual notation for a multiset's *indicator function*, which returns the multiplicity of its argument (the edge).

B.2 Graph Representation

The conceptual exposition of graphs in section B.1 already contained some hints about the mathematical/computational representation of graphs. In this section, we make explicit some mathematical and computational representations of graphs and discuss the relationships between them. Throughout this thesis, and especially in chapter 3, we present methods in terms of the mathematical represention of graphs.

B.2.1 Mathematical representation

Mathematically, a graph is usually represented by a matrix, a rectangular array of elements, denoted $\mathbf{X} = [x]_{m\times n}$, where m and n are the sizes in dimensions 1 and 2, and elements $x_{ij} \in \mathbf{X}$ $(0 < i \leq m, 0 < j \leq n$ are indexed according to a rectangular grid. Let us start from our first example in section B.1 (see Figure B.1).

Simple Graphs are usually represented by an *adjacency matrix*. An adjacency matrix $\mathbf{A}$ is the truth table of a binary relation "a is connected to b" on vertices:

$$\mathbf{A} = [\mathbf{a}]_{n\times n}, \; a_{ij} = \begin{cases} \textsc{True} & \text{if } \{v_i, v_j\} \in E \\ \textsc{False} & \text{otherwise} \end{cases} \qquad \text{(Equation 2.2)}$$

The adjacency matrix corresponding to Figure B.1 is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 \end{bmatrix} \qquad \text{(B.1)}$$

Hypergraphs Adjacency matrices cannot capture relations on a variable number of vertices well. Such general hypergraphs are usually represented by an *incidence matrix* [34]. An incidence matrix $\mathbf{I}$, with actors (vertices) on the horizontal and events (edges) on the vertical, is the truth table of a binary relation on vertices and hyperedges: "vertex j belongs

to hyperedge i"

$$\mathbf{I} = [\mathbf{i}]_{m\times n}\,,\ \mathbf{i}_{ij} = \begin{cases} \textsc{True} & \text{if } v_j \in \varepsilon_i \\ \textsc{False} & \text{otherwise} \end{cases} \qquad \text{(Equation 2.1)}$$

Incidence matrices are not restricted to general hypergraphs, but can represent simple graphs as well. Just as an example, the incidence matrix of Figure B.1 is

$$\mathbf{I} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \qquad \text{(B.2)}$$

B.2.2 Relationships between Adjacency and Incidence matrices

Intuitively, the actor-event affiliation (incidence) matrix $\mathbf{I}$ and the actor-actor or event-event adjacency matrices must be related. The relationship is in fact simple:

$$\mathbf{A}_{actors} = \mathbf{I}^{\mathsf{T}}\mathbf{I} \qquad \text{(B.3a)}$$

$$\mathbf{A}_{events} = \mathbf{I}\mathbf{I}^{\mathsf{T}} \qquad \text{(B.3b)}$$

The diagonals of $\mathbf{A}_{actors}$ and $\mathbf{A}_{events}$ encode the column and row sum (degrees of vertices and hyperedges) of $\mathbf{I}$, respectively, *i.e.* total event attendance per actor and the number of attendees per event. The incidence and adjacency matrices corresponding to the example of Figure B.2 are:

$$\mathbf{I} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix},$$

$$\mathbf{A}_{actors} = \begin{bmatrix} 2 & 2 & 1 & 1 & 1 & 1 \\ 2 & 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{A}_{events} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}.$$

Semantically, $\mathbf{A}_{actors}$ contains information about actor co-attendance at events, and $\mathbf{A}_{events}$ contains information about event co-participation by actors. Both are therefore measures of similarity.

Appendix C

SIENA

SIENA (Simulation Investigation for Empirical Network Analysis)[51, 61] allowed us to attempt to fit an Exponential Random Graph Model (ERGM) [1, 39, 52, 60] to our data. This appendix explains how SIENA was parameterised.

SIENA estimates parameters for a statistical model of a *time-evolving* network by conditioning on the change between subsequent 'snapshots' of the network. Since our data are continuous in time, the time range needed to be evenly split into a number of snapshots. In addition, SIENA only deals with binary association, so co-observation frequencies could not be incorporated directly. It was therefore necessary to find an appropriate co-observation cut-off point (see also 3) to eliminate the effect of spurious observations.

The number of snapshots and the co-observation cut-off point jointly affect the number of interactions that are retained, and the amount of change between snapshots. Creating too many snapshots will lead to low density (see 2.5) and large change between snapshots, and few connections being retained (reaching above the co-observation cut-off point). Creating too few snapshots will also lead to relatively large change between snapshots, because actors leave and join the network (structural change). In order to find a 'sweet spot' that has relatively low change between snapshots and retains relatively many connections, we computed average rate of change for all combinations of co-observation cut-off and number of snapshots. Because retention goes down rapidly with increasing cut-off (as discussed on page 34), we created 7 snapshots with a cut-off point of 1 (connections with only 1 co-observation are removed). The average change between snapshots was 60.9%, the lowest possible, and 56.3% of connections were retained.

Model Estimation

Reference [51] recommends to start by fitting a very simple model containing only a constant term that accounts for snapshot density (see 2.5). To this we added a term to account for differences in activity (degree) between actors, also following the recommendation of reference [51]. This basic initial model was not expected to be a good fit for the data, but rather a starting point to expand from.

SIENA simulates the formation of connections to stochastically increase the model fit. The approach is actor-based, allowing each actor to stochastically select new connections to make or old connections to break to maximise its own (local) model fit. Connections can be made using several slightly different mechanisms, but since the formation of links in our case is bilateral (undirected), and only on of them (the "pairwise compensatory model" [51]) is appropriate: a tie is formed based on how it affects the sum of the model fits of the two actors involved.

Finally, with the model pieces in place, several estimation algorithms could be used to estimate the model. The default algorithm, the Method of Moments (using default conditional estimation and score function 1 to estimate derivatives; see reference [51]), was chosen because the alternatives would be impractically slow on a network of this size. Other parameters of the algorithm were kept at their default values, except the number of phase 2 subphases (set to 8 instead of 4) and the number of phase 3 iterations (set to $4,000$ instead of $1,000$), to increase precision and reliability [51].

Failure to Converge

Even the above simple model failed to converge. Reference [51] gives several situations in which this may occur, and for each we have indicated if and why they could or could not apply to our data and set-up.

- Misspecification of the data.
  The data specification was verified several times.

- On rare occasions convergence may be poor due to 'bad' random initial values.
  The model estimation was attempted four times, decreasing the likelihood of the already rare event.

- The model is not appropriate for the data.
  This is the most likely case.

- The snapshots are inappropriately chosen.
  We believe that the procedure described above has generated the most appropriate snapshots that could be generate from these data.

- Too many weak effects are included.
  This is not the case in our simple model.

- One or more terms are collinear.
  This is not the case in our simple model.

- A term with a large but poorly-determined parameter is included in the model.
  Usually these are inappropriate parameters (reciprocity in an undirected network, for example), but neither of our terms is a candidate.

Bibliography

[1] Carolyn J. Anderson, Stanley Wasserman, and Bradley Crouch. A p^* primer: logit models for social networks. *Social Networks*, 21:37–66, 1999. 55, 85

[2] Laurens Bakker, Warren Hare, Hassan Khosravi, and Bojan Ramadanovic. A social network model of investment behaviour in the stock market. *Physica A*, 389(6):1223–1229, March 2010.

[3] Vladimir Batagelj and Andrej Mrvar. Pajek—program for large network analysis. *Connections*, 21(2):47–58, 1998. 35

[4] Lisa M. Batista. Season summary report—cold wet weather mat program. Technical Report December 1st 2007–March, 31st 2008, Hope for Freedom Society, 2008. 20

[5] BC Housing. Community partnership initiatives. on line, April 2011. 1

[6] Abraham Bookstein. The bibliometric distributions. *The Library Quarterly*, 46(4):416–423, October 1976. 13

[7] Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Görke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2):172–188, Feb 2008. 30, 31, 70

[8] Andreas Brandstädt, Van Bang Le, and Jeremy P. Spinrad. *Graph Classes—A Survey*. SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics, 1999. 79

[9] Statistics Canada. British columbia municipal census populations, 1921–2006. on-line. accessed February 1, 2010. 5

[10] Rhonda Chaytor. Recent advances in privacy preserving data publishing. MoCSSy Graduate Student Seminar Series, October 2009. 7

[11] City of Vancouver. Initial homeless count shows drop in street homelessness; overall homelessness increasing. press release, April 2010. http://vancouver.ca/mediaroom/news/detail.htm?row=54&date=2010-04-08. 1

[12] Aaron Clauset, Mark E.J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004. 31, 70

[13] Sam Cooper. Homeless numbers up, but more shelters mean fewer on streets—robertson to push new premier for more affordable housing. newspaper article, The Province, March 2011. 1

[14] Moniek Coumans, Maarten Cruyff, Peter van der Heijden, Hans Schmeets, and Judith Wolf. Dakloos in nederland. In *Bevolkingstrends*. Centraal Bureau voor de Statistiek, 2010. in Dutch. 1

[15] Havi Echenberg and Hilary Jensen. Defining and enumerating homelessness in canada. Technical Report PRB 08-30E, Parliamentary Information and Research Service, Library of Parliament, Canada, 2008. 2

[16] Susan T. Ennett, Susan L. Bailey, and E. Belle Federman. Social network characteristics associated with risky behaviors among runaway and homeless youth. *Journal of Health and Social Behavior*, 40(1):63–78, 1999. 3

[17] Jose Espineira. Evaluating the extent of rough sleeping. Technical report, Department for Communities and Local Government of the U.K. government, 2010. 1

[18] C. Fothergill-Payne and T. O'Halloran. Implementation review: Supporting communities partnership initiative (scpi). Technical report, Internal Audit and Risk Management Services, Human Resources and Skills Development Canada, 2001. 1

[19] Linton C. Freeman. Centrality in social networks. *Centrality in Social Networks: I. Conceptual Clarification*, 1(2):215–239, 1978–1979. 26

[20] M. Girvan and Mark E.J. Newman. Community structure in social and biological networks. *Proc Nat Acad Sci*, 99:7821–7826, 2002. 26

[21] Glasgow Homelessness Network. Where will they go?—what homeless people in glasgow said about homelessness, hostels and homelessness services. Technical report, 2003. 2

[22] Mark Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, May 1973. 34

[23] Mark Granovetter. The strength of weak ties: a network theory revisited. *Sociological Theory*, 1:201–233, 1983. 34

[24] Robert A. Hanneman and Mark Riddle. *Introduction to social network methods*. University of California, Riverside, 2005. Published on line at http://faculty.ucr.edu/~hanneman/. 79

[25] HRDC. The government of canadas homelessness initiative—supporting community partnerships initiative. Community guide, Human Resources Development Canada, August 2000. 1

[26] HRDC. Evaluation of the national homelessness initiative: Implementation and early outcomes of the hrdc-based components. Technical report, Human Resources Development Canada, March 2003. 1

[27] HRSDC. The homelessness partnering strategy—program update: Renewal of the homelessness partnering strategy. press release, November 2010. http://www.hrsdc.gc.ca/eng/homelessness/index.shtml, retrieved April 8, 2011. 1

[28] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985. 63

[29] Andrew Ilachinski. *Cellular Automata—A Discrete Universe*. World Scientific, 2001. 79

[30] Paul Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Socit Vaudoise des Sciences Naturelles*, 44:223–270, 1908. Cited in reference [49]. 62

[31] Jim Woodward and Associates Inc., Eberle Planning and Research, Deborah Kraus Consulting, and SPARC BC. From shelter to home... greater vancouver shelter strategy 2006–2015, May 2006. 1

[32] Kurt D. Johnson, Les B. Whitbeck, and Dan R. Hoyt. Predictors of social network composition among homeless and runaway adolescents. *Journal of Adolescence*, 28(2):231–248, 2005. 3

[33] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7–15, April 1989. 25, 35

[34] David Knoke and Song Yang. *Social Network Analysis*. Quantitative Applications in the Social Sciences. SAGE Publications, 2 edition, 2008. 6, 26, 34, 79, 82

[35] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: A comparative analysis. *Physical Review E*, 80(5):056117, 2009. 26

[36] Gerard Lemos and Gill Goodby. A future foretold—new approaches to meeting the long-term needs of single homeless people. Technical report, Crisis, 1999. 2

[37] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001. 52

[38] Marcelo A. Montemurro. Beyond the zipfmandelbrot law in quantitative linguistics. *Physica A*, 300:567–578, 2001. 13

[39] Martina Morris, Mark S. Handcock, and David R. Hunter. Specification of exponential-family random graph models: Terms and computational aspects. *Journal of Statistical Software*, 24(4):1548—7660, 2008. 55, 85

[40] Peter J. Mucha, Thomas Richardson, Kevin Macon, Mason A. Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010. 70

[41] Mark E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004. 29, 59

[42] Mark E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006. 29

[43] Mark E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the USA*, 103(23):8577–8582, June 2006. 30, 31, 33, 50, 61, 70

[44] Interagency Council on Homelessness. Opening doors—federal strategic plan to prevent and end homelessness. Technical report, Interagency Council on Homelessness of the U.S.A. government, 2010. 1

[45] Karl Pearson. Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. [abstract]. In *Royal Society Proceedings*, volume 59, page 69–71, 1895–1986. 17

[46] Mason A. Porter, Jukka-Pekka Onnela, and Peter J. Mucha. Communities in networks. *Notices of the American Mathematical Society*, 56(9):1082–1097, October 2009. 29

[47] Geoffrey Randall and Susan Brown. Prevention is better than cure. Technical report, Crisis, 1999. 2, 3, 70

[48] Ray Reagans. Comment on "catching the network science bug" by david l. alderson. online commentary, October–November 2008. Vol. 56, No. 5. 2

[49] Raimundo Real and Juan M. Vargas. The probabilistic basis of jaccard's index of similarity. *Systematic Biology*, 45(3):380–385, 1996. 62, 90

[50] Dennis R. Ridley. Zipf's law in transcribed speech. *Psychological Research*, 44:97–103, 1982. 13

[51] Ruth M. Ripley and Tom A.B. Snijders. Manual for siena version 4.0 (provisional version, february 14, 2010). Technical report, University of Oxford, Department of Statistics; Nuffield College, Oxford, 2010. http://www.stats.ox.ac.uk/siena/. 55, 85, 86

[52] Garry L. Robins, Tom A.B. Snijders, Peng Wang, Mark S. Handcock, and Philippa E. Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29:192–215, 2007. 55, 85

[53] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 41(1):59–66, 1988. 17

[54] Stacy Rowe and Jennifer Wolch. Social networks in time and space: Homeless women in skid row, los angeles. *Annals of the Association of American Geographers*, 80(2):184–204, 1990. 3

[55] RSCH. Vision. on line. http://www.metrovancouver.org/planning/homelessness. 1

[56] RSCH. Homelessness in metro vancouver: A comparative community profile. Technical Report 004909945, The Greater Vancouver Regional Steering Committee on Homelessness, Burnaby, B.C., Canada, March 2010. 1

[57] S.L. Satel, S.M. Southwick, and F.H. Gawin. Clinical features of cocaine-induced paranoia. *American Journal of Psychiatry*, 148:495–498, 1991. 10

[58] Henry Scheffé. *The Analysis of Variance*. Wiley and Sons, 1959. 19

[59] Scottish Homelessness Task Force. Helping homeless people. Technical report, Ministry for Social Justice of the Scottish Executive, 2002. 1, 2

[60] Tom A.B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36:99–153, 2006. 55, 85

[61] Tom A.B. Snijders and Marijtje A.J. van Duijn. *Simulating Social Phenomena*, volume 456 of *Lecture Notes in Economics and Mathematical Systems*, chapter Simulation for statistical inference in dynamic network models, page 493–512. Springer, Berlin, 1997. 55, 85

[62] Lesley Stenhouse. Social networks—why are they important to homeless people? Technical report, Shelter, 2005. 2

[63] Mohammad Tayebi, Uwe Glaesser, Laurens Bakker, and Vahid Dabbaghian-Abdoly. Locating central actors in co-offending networks. In *Proceedings of the 2011 International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, 2011. 13, 26

[64] Rob C. Thiessen. Report on the homeless in tri-cities. Technical report, The Hope for Freedom Society, April–September 2006. 5, 6

[65] Rob C. Thiessen. 2nd report on the homeless in tri-cities. Technical report, The Hope for Freedom Society, October–March 2007. 5, 6

[66] Rob C. Thiessen, 2009–2010. personal communication. 3, 5, 10, 20, 70

[67] Jennifer B. Unger, Michele D. Kipke, Thomas R. Simon, Christine J. Johnson, Susanne B. Montgomery, and Ellen Iverson. Stress, coping, and social support among homeless youth. *Journal of Adolescent Research*, 13(2):134–157, April 1998. 3, 70

[68] Krisztina Vásárhelyi. Impact of hiv testing strategies on characteristics of the undiagnosed population. Presentation at the 6th IRMACS Day, April 2011. personal communication. 3

[69] Duncan J. Watts. *Six Degrees—The Science of a Connected Age*. W.W. Norton & Company, 2003. 27

[70] Stephen Wolfram. *A New Kind of Science*. Wolfram Media, 2001. 79

[71] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, winter 1977. 30