Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

"Avaliação da Técnica de Amostragem "Respondent-Driven Sampling" na Estimação de Prevalências de Doenças Transmissíveis em Populações Organizadas em Redes Complexas"

por

Elizabeth Maciel de Albuquerque

Dissertação apresentada com vistas à obtenção do título de Mestre em Ciências na área de Saúde Pública.

Orientadora principal: Prof.ª Dr.ª Cláudia Torres Codeço
Segundo orientador: Prof. Dr. Francisco Inácio Pinkusfeld Monteiro Bastos

Rio de Janeiro, julho de 2009.

Ministério da Saúde

FIOCRUZ
Fundação Oswaldo Cruz

ESCOLA NACIONAL DE SAÚDE PÚBLICA
SERGIO AROUCA
ENSP

Esta dissertação, intitulada

"Avaliação da Técnica de Amostragem "Respondent-Driven Sampling" na Estimação de Prevalências de Doenças Transmissíveis em Populações Organizadas em Redes Complexas"

apresentada por

Elizabeth Maciel de Albuquerque

foi avaliada pela Banca Examinadora composta pelos seguintes membros:

Prof.ª Dr.ª Maeve Brito de Mello

Prof. Dr. Oswaldo Gonçalves Cruz

Prof.ª Dr.ª Cláudia Torres Codeço – Orientadora principal

Dissertação defendida e aprovada em 09 de julho de 2009.

AGRADECIMENTOS

A Deus, em primeiro lugar. Pois quando não se tem fé, nada se consegue.

Aos meus pais, José Augusto e Denise, aos meus padrinhos, Milton, Sônia e Arcília, e aos meus irmãos, Pedro e Lucas. Talvez até agora vocês não entendam bem o que é RDS com essas tais sementes que geram filhos e por aí vai... Mas obrigada por estarem sempre lá pra me ouvir, pra me acalmar, e pra "puxar a minha orelha" quando era necessário.

Aos meus orientadores Cláudia Torres Codeço e Francisco Inácio Bastos. Esse trabalho não teria acontecido se vocês não estivessem comigo. Obrigada por terem respeitado "meus momentos", principalmente "meus sumiços". Agradeço também por estarem sempre *online* para me atender, e pela motivação contínua. Uma vez me disseram que, no mestrado, ou você gosta do tema, ou dos orientadores. Eu não concordo, e me sinto muito feliz por ter conseguido conciliar as duas coisas. Foi um trabalho muito prazeroso de ser feito!

Quero fazer um agradecimento especial a Maeve Brito de Mello. Esse trabalho também não aconteceria sem a sua participação. Obrigada por todas as conversas, desde 2006, quando você me apresentou o RDS e me encorajou a mudar o rumo da minha pós-graduação. Obrigada por todos os conselhos, por me ouvir e, principalmente, obrigada por sempre ter acreditado que eu era capaz.

Aos meus amigos e outras pessoas especiais que fazem parte da minha vida. Estou pensando em cada um de vocês nesse momento, e se fosse fazer agradecimentos individuais, esse trabalho teria dois volumes. Afinal, quem de vocês não me ouviu falar da dissertação? Quem não teclou comigo na Internet enquanto eu estava "buscando inspiração"? Ou quando estava só enrolando mesmo? Quem não recebeu o recado "desculpa, tenho que escrever minha dissertação" como resposta aos convites para chopps e baladinhas? Obrigada por estarem sempre presentes e sempre me incentivando! Cada um de vocês teve uma importância única para que esse trabalho fosse concluído. Desde o início, com as felicitações por entrar no mestrado... Até agora, na reta final, me dando força pra encarar essa ansiedade enlouquecedora e compreendendo minhas chatices decorrentes da TPM (Tensão Pré-Mestre). Vocês são fundamentais!

A todas as pessoas com quem trabalho. Obrigada pelo apoio, ainda mais nessa reta final. Agradeço também por permitirem que minhas infindáveis simulações rodassem ao longo de dias, noites e fins de semana nos seus computadores. Não tenho dúvida de que sem essa ajuda, eu não terminaria essa etapa em tempo.

Á CAPES, pelo apoio financeiro ao longo desses anos.

RESUMO

Diversos fatores podem dificultar a caracterização acurada do perfil de uma população por amostragem. Se a característica que define a população é de difícil observação – seja porque exige testes caros para detecção ou porque é uma característica de comportamento ilegal ou estigmatizado que dificulta a identificação, torna-se praticamente impossível aplicar os métodos clássicos de amostragem, pois não se pode definir uma base de amostragem (*sampling frame*). Populações desse tipo são conhecidas como populações ocultas, ou escondidas, e alguns exemplos comumente estudados são homens que fazem sexo com homens, trabalhadores do sexo e usuários de drogas. Essa dissertação discute a técnica de amostragem conhecida como *Respondent-Driven Sampling* (RDS), originalmente proposta por Heckathorn (1997), e que vem sendo amplamente utilizada na estimação de prevalências de doenças transmissíveis em populações ocultas. Esse método pertence à família de amostragens por bola-de-neve, na qual os elementos seguintes da amostra são recrutados a partir da rede de conhecidos dos elementos já presentes na amostra, formando as cadeias de referência. Com este método, além das informações individuais, é possível estudar também as relações entre os indivíduos.

O recrutamento por bola de neve não gera uma amostra aleatória, e está sujeito às propriedades das redes sociais das populações em estudo, que deve mudar de lugar para lugar e potencialmente influenciar as medidas de prevalência geradas. As redes sociais são estruturas complexas, e compreender como que a amostragem RDS é influenciada por estas estruturas é um dos objetivos dessa dissertação. Além disso, se o interesse de um estudo epidemiológico é estimar a prevalência de uma doença transmissível, há de se considerar que muitas vezes a própria rede social pode estar correlacionada com as redes de transmissão, gerando potenciais dependências entre o processo de amostragem e a distribuição da variável desfecho.

Essa dissertação teve por objetivo avaliar estimativas de prevalência geradas a partir de amostras obtidas com a utilização da metodologia RDS, considerando estruturas populacionais complexas, ou seja, populações com estruturas distintas de ligação entre os indivíduos e de disseminação de doenças. Para isso, foram realizados experimentos de simulação combinando quatro modelos geradores de redes sociais e quatro modelos de distribuição de casos infectados na população. Para cada uma, foram obtidas amostras utilizando RDS e as respectivas prevalências foram estimadas.

Com os resultados encontrados, foi possível realizar uma avaliação tanto do RDS como forma de recrutamento, como o modelo proposto por Heckathorn (2002) para a ponderação e estimação de prevalências. Basicamente, três aspectos foram considerados nessa avaliação: 1. o tempo necessário para concluir a amostragem, 2. a precisão das estimativas obtidas, independente da ponderação, e 3. o método de ponderação. De forma geral, o método apresentou bons resultados sob esses três aspectos, refletindo a possibilidade de sua utilização, ainda que exigindo cautela. Os achados apresentam-se limitados, pois são escassos os trabalhos que abordem essa metodologia e que permitam estabelecer comparações. Espera-se, no entanto, despertar o interesse para que outros trabalhos nessa linha sejam desenvolvidos.

Palavras-chave: Amostragem, *Respondent-Driven Sampling*, estimação de prevalências, simulação, modelos de redes aleatórias.

ABSTRACT

Several factors may hamper the accurate characterization of a population. If the defining feature of the population is difficult to apply - either because it requires expensive tests for detection or because it is a stigmatized or illegal behavior that hinders the identification, it is virtually impossible to apply traditional methods for sampling, because sampling frame cannot be define. The latter are called “hidden populations”, and some examples are men who have sex with men, sexual workers and drug users. This dissertation focus on Respondent-Driven Sampling (RDS), a sampling method originally proposed by Heckathorn (1997), which has been widely used to estimate the prevalence of infectious diseases in hidden populations. RDS is a snowball sampling method, in which new elements for the sample are recruited from the network of the elements already present in the sample, forming reference chains. With this method, besides individual informations, it is also possible to study the relationships between individuals.

Snowball sampling does not generate random samples, and its properties are likely to depend on the properties of the social networks underlying the recruitment process, which may change from place to place and potentially influence the measures of prevalence generated. Social networks are complex structures, and understanding how the different implementations of RDS sampling is influenced by these structures is one of the objectives of this dissertation. Moreover, if the interest of an epidemiological study is to estimate the prevalence of a disease, it is should be considered that very often, social network may be correlated with the transmission networks, generating potential dependencies between the process of sampling and distribution of outcome variable.

The aim of this dissertation was to assess the behavior of prevalence estimators using RDS data in scenarios of populations organized in complex structures, i.e. Combinations of social networks structures and spreading patterns. To achieve that, theoretical experiments were performed using simulation models combining four generators of social networks and four models of distribution of infected cases in the population. For each one, samples were obtained using RDS and prevalence, estimated.

Findings were used to evaluate RDS as a recruiting process itself, as well as Heckathorn’s (2002) model to estimate prevalences. Three aspects were considered in such analyses: 1. the time elapsed before obtaining the sample; 2. the accuracy of the estimates without taking in consideration the weighting strategies; and 3. the weighting strategy. Overall, RDS performed well in these three areas, showing it is a valid method to assess hidden populations, despite the fact its use should be made with the necessary caution. The interpretation of our findings was constrained by the scarcity of studies using the same methodology, what compromised the comparability of our findings. We hope, however, that our findings may foster the development of additional studies in this field.

Key words: Sampling, Respondent-Driven Sampling, prevalence estimation, simulation, network models.

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

GLOSSÁRIO

População oculta: Uma população é dita oculta, ou escondida, quando não existe como enumerar todos os seus membros. Em geral, seus membros são caracterizados por apresentarem comportamentos ilegais ou estigmatizados.

Rede social: Uma rede é uma estrutura social, composta por unidades individuais que estão conectadas por um ou mais meios de interdependência (idéias, amizade, contato sexual, etc.).

Vértice: Um vértice é uma unidade individual em uma rede, que pode ser representado por um elemento da população, uma organização, etc. Pela teoria dos grafos, um vértice é um elemento do gráfico, que é conectado por duas ou mais ligações.

Ligação: Forma de conexão entre dois ou mais vértices. Esses meios de interdependência podem se dar pelo compartilhamento de idéias, de relações de amizade, ou de relações sexuais, por exemplo. Pela teoria dos grafos, uma ligação é a forma de conexão entre dois vértices.

Grau: Número de pessoas que uma pessoa conhece diretamente, ou o número de ligações de um vértice.

Sementes: São os indivíduos que iniciam um processo de amostragem utilizando cadeias de referência, e os quais são escolhidos de forma não aleatória.

Filhos/Frutos: Os primeiros filhos em um processo de recrutamento são os indivíduos recrutados pelas sementes. Da mesma forma, os indivíduos recrutados por eles, serão também chamados de filhos até que a amostra esteja completa.

Onda de recrutamento: As ondas de recrutamento são formadas a medida que novas pessoas entram na amostra. Assim, as sementes pertencem a onda 0 (zero) do processo de recrutamento. Os filhos gerados por essas sementes pertencerão a primeira onda do recrutamento e assim por diante, até que a amostra esteja completa.

1. Introdução

Uma população é o conjunto de todos os elementos ou resultados de determinada investigação (Bussab & Morettin, 2007). Ao aplicar essa definição à epidemiologia, as populações de maior interesse são humanas, constituídas por um conjunto de pessoas que tenham pelo menos uma característica em comum. Como uma das questões centrais da epidemiologia consiste em quantificar a ocorrência de doenças em populações (Rothman & Greenland, 1998), contrair ou não uma doença, e estar exposto ou não a um determinado fator em comum, definem conjuntos de características largamente utilizadas nesse contexto. Dessa forma, tem-se, de um lado, estudos buscando conhecer e descrever algumas dessas características – como morbidade, mortalidade e seus determinantes – e, de outro, o trabalho da vigilância epidemiológica, onde um dos principais objetivos é a caracterização do perfil de epidemias locais, visando subsidiar intervenções mais efetivas e permitir seu monitoramento (Magnani *et al.*, 2005).

Diversos fatores podem dificultar a caracterização acurada do perfil de uma população. Por exemplo, o seu tamanho, se muito grande, pode tornar inviavelmente elevado o custo necessário para investigar todos os seus indivíduos. Além disso, se a característica que define a população é de difícil observação – seja porque exige testes caros para detecção ou porque é uma característica de comportamento ilegal ou estigmatizado – a delimitação de quem pertence ou não à população pode tornar-se inviável ou impossível. Populações desse último grupo são conhecidas como populações ocultas, ou escondidas (Heckatorn, 1997), e alguns exemplos comumente estudados são homens que fazem sexo com homens (HSH), trabalhadores do sexo (TS) e usuários de drogas injetáveis (UDI).

Em todos esses casos, por não ser trivial obter informações da população inteira, técnicas de amostragem são largamente utilizadas. Uma amostra é um subconjunto de uma população, e as técnicas de amostragem são os meios pelos quais essas amostras são obtidas, de forma que elas representem a população corretamente, respeitando, por exemplo, as proporções reais de suas características. A técnica mais apropriada para cada situação é determinada pelo interesse do estudo, assim como pelas características da população estudada e os recursos disponíveis. Em geral, com exceção dos censos e inquéritos exaustivos de um dado segmento (não oculto), todas as demais abordagens em epidemiologia têm por base amostras, sejam estas representativas (por exemplo,

amostragem aleatória simples ou estratificada) ou de conveniência (por exemplo, casuísticas clínicas ou estudos de coorte), sendo as de conveniência mais frequentes (Semaan *et al*., 2002).

Essa dissertação discute a técnica de amostragem conhecida como *Respondent-Driven Sampling* (RDS), originalmente proposta por Heckathorn (1997), e que vem sendo amplamente utilizada na estimação de prevalências de infecções/doenças transmissíveis em populações ocultas. Alguns estudos que utilizaram essa metodologia são Heckathorn *et al*. (2002), Ramirez-Valles *et al*. (2005), Robinson *et al*. (200.6) e Wattana *et al*. (2007). Como será explicado em detalhes posteriormente, o método RDS faz parte da família de métodos de amostragem "bola de neve", que utilizam cadeias de referência para o recrutamento. Diferente das técnicas tradicionais de amostragem, que buscam a independência entre os elementos da amostra, esse tipo de técnica faz uso justamente das relações entre as pessoas. Em poucas palavras, no processo de recrutamento RDS cada participante da amostra recebe um número limitado (previamente definido pelos pesquisadores) de convites e é estimulado a trazer os próximos participantes que farão parte da amostra, através da entrega desses convites. Em relação ao método de bola de neve tradicional, o RDS visa minimizar o fenômeno de "clonalidade" das ondas de recrutados. Ou seja, como, em geral, o método tradicional não restringe as sucessivas nominações, é possível que um único recrutador (ou uns poucos recrutadores), que a literatura em língua inglesa habitualmente denomina *super-recruiter(s)*, imponha(m) um padrão "clonal" a uma rede com características heterogêneas, ou seja, um padrão que antes repete as características do(s) super-recrutador(es) do que as da rede em si. Uma forma extrema de clonalidade é o recrutamento por mais de uma vez de uma mesma pessoa, mas isso pode ser controlado mediante procedimentos que impedem o comparecimento múltiplo de um mesmo indivíduo (utilizando desde prova de identidade a medidas antropométricas), mas a clonalidade não se limita ao comparecimento por várias vezes (tributário da má fé ou não) de um dado indivíduo, podendo, na verdade, ser fruto da seletividade das redes de um ou mais de um super-recrutador, que basicamente interage com pessoas com características semelhantes as dele (ou dela) mesmo/a (Díaz et al., 1992).

As cadeias de referência surgem do processo de pessoas recrutarem outras pessoas dentre seus conhecidos. Esse conjunto de relações de conhecimento e amizade entre elas é denominado "rede social" pela sociologia, e, assim, o método de bola de neve pode ser visto como um método de percorrer caminhos nesta rede social. Contudo,

como será apresentado a seguir, essas relações não são necessariamente as mais relevantes para os estudos epidemiológicos (Morris, 2004).

O conceito de rede também tem sido utilizado na epidemiologia, em contextos um pouco distintos do da sociologia. De um lado, há a necessidade de compreender e descrever o processo de encontros e interação entre pessoas, ou seja, as suas redes de conhecidos, ou redes sociais, através das quais comportamentos de risco ou de proteção podem se propagar. E, por outro lado, há as estruturas por onde as infecções/doenças são propagadas, ou seja, as redes de transmissão de infecções/doenças, cujas ligações dependem do modo de transmissão de cada infecção/doença. Ambos os conceitos foram definidos por Luke & Harris (2007) como relevantes à saúde pública e são abordados nessa dissertação.

O estudo de redes mostra que populações reais interagem de forma não aleatória e dinâmica e que, portanto, os fenômenos de disseminação, por exemplo, de inovações, informações e doenças, não se dão num espaço uniforme ou direcional. Nesse sentido, as redes são vistas como objetos dinâmicos, que estão mudando o tempo todo, dependendo do comportamento prévio e atual de seus componentes e da natureza das suas interações.

As múltiplas redes que formam o tecido social podem estar correlacionadas ou não, e provém daí a denominação de populações organizadas em redes complexas (Watts, 2003 e Barabási & Albert, 1999). Ao assumir que uma população está estruturada sob a forma de grupos, ou seja, de redes com/em interação, correlações começam a surgir entre o evento individual ─ "ter a doença" ─ e as características que identificam os próprios grupos, como idade e comportamento(s). Por exemplo, pessoas utilizando o mesmo ônibus, ou que se encontram em *shoppings*, podem ter grande importância na rede de transmissão de doenças respiratórias. Da mesma forma, encontros sexuais entre pessoas que pouco se conhecem ou compartilhamento de seringas em festas e outros eventos, que são comuns em algumas populações ocultas, também são importantes nas redes de transmissão de infecções/doenças sexualmente transmissíveis (ISTs/DSTs) (Friedman *et al.*, 2007). Assim, para compreender a dinâmica de transmissão de uma infecção/doença, faz-se necessário proceder à caracterização dos agrupamentos, o que apesar de ser muito importante, pode ser extremamente difícil, quando fala-se por exemplo de populações ocultas, devido ao caráter casual ou sigiloso dos agrupamentos, assim como a complexidade das suas estruturas (Wallinga *et al.*, 1999).

Finalmente, por ser relevante para os estudos epidemiológicos o conhecimento do perfil de determinadas populações, como as ocultas, por exemplo, e a necessidade de se ter boas estimativas sobre a prevalência de determinadas doenças, como as DSTs, essa dissertação tem por objetivo avaliar estimativas de prevalências obtidas a partir de amostras que utilizam a técnica RDS, considerando populações organizadas em redes complexas. Além disso, na literatura, alguns estimadores ponderados são apresentados para dados obtidos a partir de amostras que contaram com a metodologia RDS (Heckathorn 1997, Heckathorn 2002, Volz & Heckathorn, 2008). Se, por um lado, o modelo teórico explica que as estimativas são aproximadas e tem-se que suas propriedades ainda não foram exaustivamente avaliadas, por outro, todo o procedimento envolvendo RDS consiste num complexo processo estocástico. Por essa razão, emerge a decisão de se utilizar modelos matemáticos e técnicas de simulação computacional para avaliar as estimativas de prevalência geradas utilizando a técnica RDS. Considerando ainda que as amostras obtidas nessa dissertação sejam tributárias de diversas estruturas, tanto de redes sociais, como de redes de transmissão de doenças, essa avaliação não poderia ser realizada de outra forma. Ainda assim, na tentativa de representar as populações simuladas de forma mais próxima a populações reais, dados empíricos foram utilizados. Esses dados se referem ao projeto “Semear Saúde”, realizado em Campinas entre 2005 e 2006, cujos objetivos compreendiam desenhar o perfil da população HSH residente na região metropolitana de Campinas e estimar a prevalência de HIV/Aids nessa população, utilizando a amostragem RDS (Mello *et al.*, 2008).

O capítulo 2 apresenta uma revisão sobre modelos de redes sociais e redes de transmissão de doenças, bem como traz uma explanação detalhada sobre a metodologia RDS. No capítulo 3 são apresentados os objetivos dessa dissertação. No capítulo 4 serão explicados os modelos gerados e os parâmetros utilizados. Por fim, o capítulo 5 apresenta os resultados encontrados e o capítulo 6 traz uma discussão acerca das lições aprendidas e trabalhos que podem ser desenvolvidos no futuro.

2. Revisão da literatura

2.1. Modelos de redes sociais e redes de transmissão de doenças

A maioria dos fundamentos teóricos da epidemiologia clássica de doenças infecciosas se baseia no pressuposto de que os contatos entre os membros de uma população são aleatórios, mas, na prática, cada indivíduo tem um grupo de contatos mais próximos e com características mais semelhantes, fato que pode influenciar, por exemplo, a dinâmica de transmissão de determinada doença. Por isso, os estudos de redes são muito importantes para os estudos epidemiológicos, conferindo-lhes um caráter não apenas atual, como mais próximo do mundo real, onde os fenômenos que a epidemiologia analisa ocorrem (Keeling & Eames, 2005).

As redes de contato são as estruturas que descrevem o padrão de interação entre as pessoas (quem encontra quem). Contatos podem ser definidos de múltiplas formas (contatos físicos íntimos ou não íntimos, duradouros ou não, contatos indiretos, etc.). Cada infecção/doença terá sua definição mais apropriada de contato. Cada forma de contato define uma rede: rede de parceiros sexuais, rede de usuários de drogas, redes de homens que fazem sexo com homens, redes de amizades, etc. Uma maneira de classificar redes é a partir da distribuição dos graus individuais, onde o grau de uma pessoa é definido como o número de pessoas com quem ela tem contato, ou o número de pessoas as quais ela está diretamente conectada (Salganik & Heckathorn, 2004). A distribuição de frequência dos graus observados numa rede é um descritor da topologia desta rede. Embora a distribuição de graus possa seguir diferentes formas, a revisão apresentada nessa dissertação refere-se às redes aleatórias, onde os graus podem seguir uma distribuição de Poisson, uma distribuição exponencial e uma distribuição que segue uma lei de potência (livres de escala) (Barabási & Albert, 1999).

Antes, porém, faz-se necessário introduzir dois conceitos que serão muito utilizados nos próximos parágrafos, pois a teoria apresentada sobre essas estruturas se baseia na teoria de grafos. Na teoria dos grafos, uma rede é representada por um grafo, composto por um conjunto de vértices e ligações entre vértices (arestas). Vértice, no contexto dessa dissertação se refere a cada elemento da população, ou seja, cada indivíduo é reconhecido como um vértice. O outro conceito é a formalização das ligações, que representam as conexões entre as pessoas. Dessa forma, no contexto de uma rede social, dizer que dois vértices estão conectados por uma ligação significa dizer que duas pessoas se conhecem (Scott, 2000).

As redes aleatórias são caracterizadas pela irrelevância da posição dos indivíduos, em termos matemáticos, ou seja, são redes onde todos os indivíduos têm a mesma chance de a ela pertencer (Stumpf & Wiuf, 2005). Com isso, esse tipo de estrutura de rede não gera grupos (partes da rede que ficam mais conectadas entre si do que com o entorno) e o número esperado de ligações é o mesmo para cada pessoa, com probabilidade **p** de ocorrência e independente da sua posição específica (Bollobás, 2001). Em uma das suas formas de construção, as conexões entre os indivíduos se dão de maneira aleatória. Suponha uma população formada por N vértices e onde cada vértice tenha uma média de *z* ligações. Assim, segundo Newman *et al.* (2001), os graus dos vértices da rede terá a seguinte distribuição de probabilidade:

$$p_k = \left(\frac{N}{k} \right) p^k (1-p)^{N-k} \qquad (1)$$

onde **k** é uma constante maior do que zero. Por outro lado, **p** pode ser obtido tal que **p** = *z*/(N-1) = *z*/N, se N for um número grande e assim, a equação acima pode ser aproximada para

$$p_k \approx \frac{z^k e^{-z}}{k!} \qquad (2)$$

ou seja, a distribuição dos graus dos elementos da população segue uma distribuição de Poisson, com parâmetro z.

As redes aleatórias com distribuição de grau seguindo uma distribuição de Poisson não constituem bons modelos para redes sociais. Como já foi dito, é sabido que, em geral, conjuntos de características como idade, religião, local de moradia, etc., são relevantes no estabelecimento das relações entre pessoas, e esse conjunto de características não é considerado nos modelos aleatórios. As distribuições dos graus das redes de contato em geral são melhor representadas por modelos livres de escala e/ou modelos exponenciais. Estes modelos são capazes de representar a tendência destas distribuições de graus, que tendem a ter poucos vértices com muitas ligações e muitos vértices com poucas ligações.

Barabási & Albert (1999) propuseram um modelo de criação de redes que tenta imitar o modo pelo qual as redes de contato são formadas, o modelo livre de escala. Nele, o processo de construção da rede se dá pela introdução de um vértice de cada vez. À medida que um novo vértice é introduzido, ele é conectado a outro, de modo que haja uma probabilidade maior desta ligação ser estabelecida com vértices que apresentam

maior grau. A distribuição de probabilidade de graus resultante deste processo de formação segue uma Lei de potência:

$$p_k \approx k^{-\gamma} \tag{3}$$

com γ sendo um parâmetro estabelecido a partir de um ajuste dos dados da população de referência. Essa classe de modelos tem apresentado bons resultados na descrição de redes não biológicas, reproduzindo bem a interconexão entre os elementos e estimando corretamente a amplitude do parâmetro γ (Laird & Jensen, 2006).

Finalmente, tem-se o modelo exponencial, onde a rede também é construída com a inclusão de um vértice de cada vez. Nesses modelos, as ligações entre seus elementos podem se dar de duas formas. Uma delas leva em consideração a distribuição dos graus individuais, assim como no modelo livre de escala, ou seja, há uma probabilidade maior da ligação ser estabelecida com vértices que apresentem maior grau. Se isso não acontece, a ligação ocorre com base numa probabilidade **p**, independente para cada ligação, que é gerada pela combinação de uma série de características. Técnicas como simulação de Monte Carlo para cadeias Markovianas podem ser empregadas para a geração dessas redes. Sobre a distribuição dos graus dos elementos gerados por esse modelo, como o próprio nome sugere, ela segue uma distribuição exponencial, tal que

$$p_k = \left(1 - e^{-1/\lambda}\right) e^{-k/\lambda} \tag{4}$$

onde k é uma constante maior do que zero e λ é um parâmetro estabelecido a partir de um ajuste de dados (Volz, 2004). Comparando os dois últimos modelos, tem-se que, à medida que os valores dos graus vão aumentando, a distribuição de graus gerada pelo modelo exponencial apresenta um decaimento mais acentuado do que aquele potencialmente gerado pela lei de potência (Newman *et al.*, 2002).

Algumas redes de contato podem servir também de canais para transmissão de agentes patogênicos, definindo dessa forma redes de transmissão de doenças. Modelos para estudar as dinâmicas de transmissão em redes estão em expansão na literatura (Meyer set *al.*, 2006 e Luke & Harris, 2007) sendo aplicados para modelar a propagação de doenças infecciosas, levando em consideração as estruturas de ligação entre pessoas em uma população. Os primeiros modelos desenvolvidos para analisar a dinâmica de doenças transmissíveis assumiam homogeneidade no risco de transmissão, ou seja, ausência de estrutura social. Dessa forma, o risco de uma pessoa se infectar dependia do

número de pessoas infectadas na população, sem se preocupar com quem eram essas pessoas e como essas interações se definiam (na verdade, não havia a preocupação em defini-las). Essa abordagem gerou duas classes de modelos que foram amplamente utilizados com sucesso, mesmo sem considerar os detalhes sobre como as infecções progrediam (Keeling & Eames, 2005). São eles os modelos SIR e SIS, onde S significa suscetível, I infectado e R recuperado, ou seja, imune (ou seja, excluído do grupo sob risco) (Ross, 1916 e Bailey, 1958). No modelo SIR, o indivíduo da população é suscetível, contrai a infecção e se recupera, não voltando a ficar suscetível (pelo menos por um extenso período de tempo). Esse modelo caracteriza bem doenças infecciosas que induzem imunidade prolongada no indivíduo, como sarampo e coqueluche. Já nos modelos SIS, depois de se infectar e se recuperar, o indivíduo volta a ser suscetível, podendo contrair a infecção novamente. Esse tipo de modelo é usado predominantemente para caracterizar doenças sexualmente transmissíveis, como sífilis e gonorréia. Vale destacar que esses modelos não podem ser usados, por exemplo, para a caracterização de doenças para as quais não existe cura (ou recuperação espontânea), como a infecção por HIV, por exemplo. Neste caso, o modelo correto seria da forma SI, ou seja, uma vez infectado, o indivíduo se mantém infectado por toda a vida.

Em 1985, Klovdahl propõe “a conceitualização da população como um conjunto de indivíduos ligados entre si formando uma grande rede”. Essa abordagem proporcionou um melhor entendimento da disseminação de doenças infecciosas. Logo no início da epidemia de Aids, por exemplo, ele assinalou que a extensão da transmissão da epidemia dependeria da estrutura das relações entre as pessoas e que a compreensão destas estruturas seria útil para a estimação do potencial epidêmico. Nesse sentido, outro estudo, realizado em Manitoba, Canadá, procedeu a um levantamento de redes sexuais e identificou 1503 componentes, com tamanhos variando entre 2 e 82 pessoas, em seis meses de estudo. Dentro desse componente maior, de 82 pessoas, foi identificado que uma delas era quem fazia a ligação entre o centro da cidade e uma área periférica (Wylie & Jolly, 2001). E, com isso, foi possível enxergar uma nova dimensão nos estudos de epidemias.

Além de analisar os padrões de disseminação das infecções, os modelos desenvolvidos para estudar as epidemias passam a incluir também informações sobre a rede social dos seus elementos, como os graus dos indivíduos e informações que caracterizem um conjunto de fatores de risco que determinam a sua chance de adquirir uma dada infecção. Segundo Meyer *et al.* (2006), modelos que consideram o tamanho

da rede de conhecidos dos indivíduos na previsão e análise da infecção podem gerar informações importantes sobre a disseminação de uma infecção. Os autores citam o estudo apresentado por Meyer *et al.* (2003), em que, para analisar a disseminação de uma infecção por Mycoplasma dentro de uma clínica, foram estudadas as redes de contato "semi-indiretas" para profissionais que atuavam na clínica e pessoas que transitavam na clínica. Essas redes de contato tinham por base não apenas os contatos diretos dos elementos analisados, mas também os contatos indiretos, como o número de relações paciente-profissional. Dessa forma, o estudo mostrou a relevância de se incluir informações sobre o número de contatos ao analisar a dinâmica de transmissão da doença.

Ainda no sentido de que incluir informações sobre os indivíduos de uma população pode ajudar a explicar padrões de transmissão de doenças, Koopman (2004) cita que para responder a perguntas como "Intervenções de controle de epidemias devem ser realizadas de forma genérica para todas as pessoas, ou direcionadas para determinados grupos?" é necessário incluir informações individuais nos modelos. No entanto, além de características sócio-demográficas, são necessárias também informações que tenham relação específica com os fatores de risco ou proteção para a infecção/doença sob análise. Dessa forma, segundo ele, é possível elaborar modelos causais mais confiáveis, que permitem responder questões específicas. Adicionalmente, é possível pensar que o conhecimento desses fatores de risco leva a pensá-los como modelos assortativos e disassortativos (observando-se que padrões de mistura puramente assortativos ou disassortativos constituem "tipos ideais", no sentido Weberiano, não existindo enquanto tais no mundo real). Os padrões assortativos ocorrem quando os indivíduos se conectam por terem alguma características em comum no mesmo sentido (*like-with-like*), ao passo que, se os indivíduos se conectam justamente devido a essa característica em comum se dar no sentido contrários (*like-with-unlike*), tem-se padrões disassortativos (Anderson, 1996).

O primeiro desafio do estudo de redes de transmissão surge, porém, do fato da metodologia estar fortemente baseada na análise de dados de redes completas, que são possíveis quando grupos pequenos são estudados. À medida que os grupos estudados não são tão pequenos, ou são difíceis de serem acessados, obter os dados completos de todas as redes de contato é praticamente impossível. Nesse sentido, para permitir que esses estudos sejam realizados, uma estratégia é a utilização de técnicas de amostragem que utilizam cadeias de referência, como será apresentado a seguir.

2.2. Amostragem por cadeia de referência

Nessa seção, serão descritas algumas técnicas de amostragem que utilizam cadeias de referência. O método mais completo seria coletar o máximo de informações sobre todos os membros da rede (*complete network design*) ou utilizar uma amostra aleatória dos participantes (*local network design*). No entanto, muitas vezes isso não é viável e o método de bola de neve se encontra no meio desse caminho, justamente por utilizar a abordagem de cadeias. Com essas técnicas é possível coletar informações em dois níveis: o primeiro são as informações sobre os indivíduos participantes da amostra, e o segundo são as informações referentes às relações entre esses participantes. Dentro das variações dessa amostragem em rede, como é denominada, o desenho é escolhido de acordo com a maneira como os participantes serão selecionados (Morris, 2004).

No método bola-de-neve, apresentado por Goodman (1961), um indivíduo é recrutado e, em seguida, indica outras pessoas de seu relacionamento para que também participem da amostra. Para isso, um número inicial de pessoas, que, preferencialmente, conhece muitos componentes da população-alvo, é selecionado. Esse grupo recebe a designação de “sementes”, por serem os primeiros indivíduos recrutados. O passo subseqüente é solicitar a essas pessoas informações acerca de outros membros da população de interesse, para, então, recrutá-los. Os próximos membros que farão parte da amostra recebem a designação de “filhos”, ou frutos, por terem sido gerados pelas sementes, e o seu recrutamento pode se dar de várias formas. Em alguns estudos, as sementes recrutam o maior número de pessoas possível; em outros, os próprios pesquisadores efetuam esse recrutamento, através de agentes que atuam em um dado campo, com conhecimento aprofundado e trânsito em uma dada comunidade (*outreach workers*). Esse procedimento é repetido algumas vezes, até que o tamanho pré-definido da amostra seja alcançado ou até que a população fique saturada (ou seja, se esgotem os membros acessíveis da mesma). A figura 2.1, a seguir, apresenta um exemplo hipotético de uma cadeia gerada por um recrutamento com essas características. Nessa figura, os círculos maiores representam as sementes, enquanto os demais são os filhos gerados por elas.

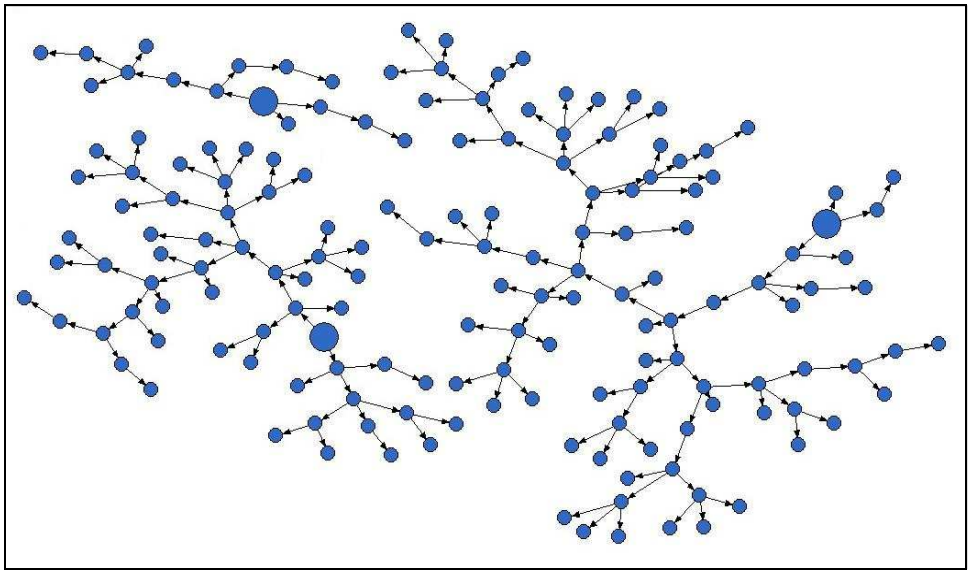

Figura 2.1. Representação hipotética de uma cadeia de referência.

Devido à sua estratégia de recrutamento, o método bola-de-neve é considerado não probabilístico, uma vez que não é possível determinar a probabilidade de seleção de cada participante na amostra. Dessa forma, não há garantia de que a amostra resultante seja não enviesada e seus resultados não podem, via de regra, ser generalizados (Semaan *et al.*, 2002). Uma vantagem dos métodos que utilizam cadeias de referência é que, em uma população oculta, é mais fácil um membro da população conhecer outro membro do que os pesquisadores identificarem os mesmos. Além disso, a amostragem por bola-de-neve pode ser muito útil em pesquisas formativas, onde o objetivo é conhecer a população estudada e/ou testar os instrumentos que serão utilizados. Por outro lado, uma limitação se refere ao fato de que as pessoas acessadas pelo método são aquelas mais visíveis na população. Em se tratando de populações ocultas, as pessoas acessadas serão aquelas que assumem determinados comportamentos e atitudes que as tornam membros dessas populações. Deve ainda ser considerado que, visando assegurar a privacidade daqueles que realmente se escondem, algumas informações sobre as pessoas conhecidas podem ser, deliberada ou involuntariamente, suprimidas.

Outro método de amostragem, conhecido como *target sampling*, foi desenvolvido na tentativa de superar algumas limitações do método de bola-de-neve. Visando reduzir possíveis vieses, o recrutamento se inicia com um mapeamento etnográfico da região e da população de interesse (Magnani *et al.*, 2005 e Heckathorn, 1997). O objetivo é delinear fronteiras geográficas que podem facilitar a realização de intervenções e descrever subgrupos da população, bem como suas redes sociais. Um exemplo desse tipo de mapeamento, apresentado num estudo de Singer *et al.* (2000),

utilizou seis métodos qualitativos distintos para examinar as diferenças de acesso a seringas esterilizadas por UDIs em três cidades, sendo um deles o mapeamento etnográfico. Segundo os autores, oito localidades foram investigadas em cada cidade onde o estudo foi realizado e esse mapeamento permitiu, entre outras coisas, a construção de um mapa das mudanças nas diferentes localidades, em termos de uso de drogas, crimes com vítimas e atuação do governo municipal, além do acesso às seringas. Em longo prazo, esse mapeamento poderá ser utilizado também para avaliar as mudanças ocorridas. Ainda assim, a conclusão dos autores foi que nenhum dos seis métodos, individualmente, é capaz de descrever a vida e o comportamento dos UDIs de forma exaustiva.

Retomando a discussão acerca da técnica de amostragem, feito o mapeamento etnográfico, a amostra é gerada, selecionando membros em cada subgrupo identificado. Nesse método, o sucesso da amostra dependerá da pesquisa etnográfica realizada anteriormente. Dessa forma, considerando um mapeamento adequado, o *targeted sampling* é considerado um bom método dentre os métodos não probabilísticos, pois inclui pessoas sob diferentes níveis de risco e provenientes de diferentes localidades. Uma de suas limitações, no entanto, é o custo elevado da pesquisa etnográfica, e o tempo necessário para sua realização, que nem sempre convergem com o da pesquisa principal.

Mais recentemente, um novo método de implementação da amostragem bola-de-neve foi proposto por Heckathorn (1997, 2002) e denominado *Respondent-Driven Sampling* (RDS). Na realidade, como será descrito a seguir, o método é bastante semelhante à bola-de-neve. No entanto, duas diferenças devem ser citadas. A primeira é que devido à forma de recrutamento, utilizando RDS é possível calcular a probabilidade de seleção de cada indivíduo. Isso faz com que o método seja considerado por alguns como um método probabilístico, embora ainda haja críticas em relação à possibilidade de generalização de seus resultados, como também será discutido a seguir. Outra diferença importante está no fato de que para dados obtidos por amostragem utilizando RDS, também foram propostos alguns modelos teóricos usados para a estimação de proporções que consideram o efeito desse desenho.

2.3. *Respondent-Driven Sampling*

O procedimento de amostragem *Respondent-Driven Sampling* se inicia com a escolha, não aleatória, de um grupo de membros da população-alvo, denominados sementes. A cada semente, é dado um número fixo de cupons, em geral três ou menos, de numeração única, que deverão ser entregues para outros membros elegíveis da população-alvo, recrutados pelo próprio participante, dentro de sua rede pessoal de conhecidos. Uma vez que esses novos indivíduos cheguem ao estudo, se eles realmente são elegíveis e desejam participar, eles passam a fazer parte da primeira onda de recrutamento e passam a ser denominados filhos das sementes que os trouxeram. O mesmo procedimento de cupons é feito com os membros da onda 1, e as pessoas trazidas por eles, nas mesmas condições, passam a integrar a segunda onda de recrutamento. A figura 2.2 ilustra esse processo, que é repetido até que o tamanho da amostra desejado seja obtido, a população-alvo se esgote ou que o tempo/recursos alocados para a pesquisa acabem. A vantagem desse processo de seleção é a redução do viés de "mascaramento", uma vez que as pessoas não precisam designar outras, mas sim convidá-las diretamente, isso faz com que essas outras pessoas passam a ter o direito de receber ou recusar o convite. Além disso, o pequeno número de cupons minimiza a influência das sementes na composição final da amostra, impedindo que se estabeleçam super-recrutadores.

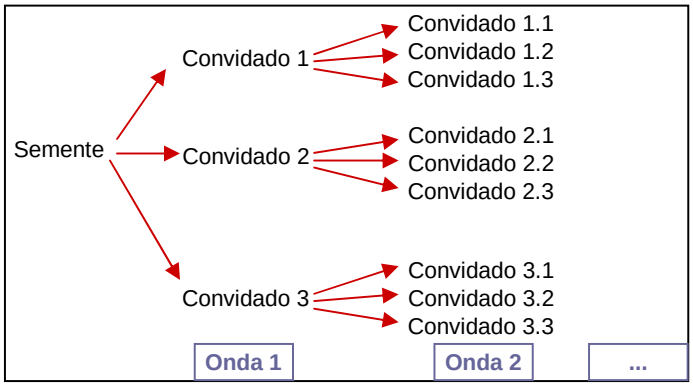

Figura 2.2. Esquema de geração da amostra com a metodologia RDS.

Uma característica importante dessa metodologia é a utilização de um sistema de duplo incentivo, que foi bastante ressaltado por Heckathorn (1997). Estes incentivos podem ser em dados em dinheiro ou em objetos de interesse da população-alvo, como ingressos para espetáculos/*show*s ou tratamentos de beleza, e variam de população para população. Dessa forma, o recrutado recebe um incentivo primário por participar do estudo e um incentivo secundário por cada participante elegível que leva ao estudo. O

objetivo desse sistema é reduzir o viés de não resposta e buscar aumentar o comprometimento dos indivíduos com o recrutamento. Heckathorn considera ainda o segundo incentivo como mais importante do que o primeiro, pois a pessoa que está convidando os elementos que farão parte da próxima onda de recrutamento pode exercer alguma pressão para que seu convidado participe do estudo.

Uma limitação do método é que, por utilizar cadeias de referência, é necessário que as pessoas da população-alvo estejam conectadas entre si, ou seja, se conheçam. Além disso, é preciso dispor de mecanismos para verificar se o participante realmente pertence à população-alvo. Embora isso não constitua exatamente uma limitação intrínseca, é preciso estar atento na hora de realizar um estudo, de modo a não enviesar a amostra, por exemplo, subestimando ou superestimando determinadas características.

Finalmente, por se tratar de um método relativamente recente e que vem sendo muito utilizado, ainda há muito que ser pesquisado, compreendido e aprimorado. Para isso, o primeiro passo é compreender o embasamento teórico do RDS, assim como o modelo proposto para estimar prevalências a partir de dados coletados com essa metodologia, assuntos que serão apresentados a seguir.

2.4 Estimativas de prevalência em dados obtidos por RDS

Com base na forma como o recrutamento é feito, Heckathorn (1997) propõe modelar o RDS como um processo estocástico markoviano, regular, de ordem 1.

Uma variável X é denominada variável aleatória quando existem x1, x2,..., xn que assumem respectivamente probabilidades - P(x1), P(x2),..., P(xn) - entre 0 e 1 e onde P(x1) + P(x2) + ... + P(xn) = 1.

Intuitivamente, se uma variável aleatória unidimensional é um número real que varia aleatoriamente, um processo estocástico é uma função que varia aleatoriamente. Mais formalmente, um processo estocástico é definido como uma família de variáveis aleatórias Y, onde Y(t) t=0,1,2,... é uma variável aleatória e t determina o estágio observado, que, em geral, é o tempo, mas no caso de RDS, representa as ondas.

Para compreender o significado de um processo markoviano é preciso compreender primeiro o que é o espaço de estados de um processo e o que é probabilidade condicional. O espaço de estados são os valores que cada variável

aleatória pode assumir, ou seja, são os estados de cada variável. No exemplo acima, x1, x2,..., xn. Considere agora dois eventos quaisquer A e B. A probabilidade condicional é definida tal que P(A| B) = P(A∩B) / (B).

Um processo markoviano pode ser finito ou infinito. Nos processos markovianos finitos, comumente citados como cadeias de Markov, o espaço de estados é finito (limitado), ou seja, n é finito e conhecido. Além disso, a probabilidade de Xn+1 depende apenas de Xn: P(Xn+1= x | Xn, Xn-1,..., X1) = P(Xn+1 = x | Xn). Intuitivamente, se o processo de escolha se der pela raça, podemos ter, por exemplo, n = 4 (branca, negra, indígena e outra). Considere-se a semente A, branca. Ela escolhe uma pessoa B, negra, para fazer parte da onda 1. Essa pessoa B deverá escolher uma pessoa C, onde a cor dessa pessoa independe da pessoa A que a escolheu. Considerando os 4 estados para raça, o esquema abaixo apresenta as probabilidades associadas a essas escolhas:

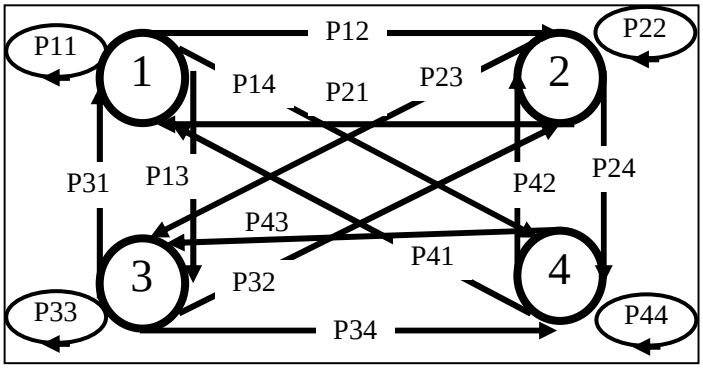

Figura 2.3. Exemplo de probabilidades de transição entre estados.

Formalmente, essas probabilidades definem uma matriz de transição de probabilidades, como descrita abaixo (considerando quatro estados):

$$\mathrm{Pij} = \mathrm{P}(\mathrm{Xn}+1 = \mathrm{j} \mid \mathrm{Xn} = \mathrm{i}), \text{ tal que } Pij = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \\ P_{41} & P_{42} & P_{43} & P_{44} \end{bmatrix}$$

Um processo markoviano tem algumas propriedades matemáticas interessantes, que são utilizadas para gerar estimadores a partir de dados amostrados por RDS. Uma delas é um importante teorema válido nesses casos, a Lei dos grandes números para cadeias regulares de Markov. Esse teorema diz que a partir de certo número de estágios,

a matriz de transição de probabilidades para processos desse tipo será a mesma para os demais estágios e será independente do estágio inicial (Kemeny & Snell, 1960). Quando isso acontece, diz-se que o processo atingiu o equilíbrio. A partir dessa informação, foi possível postular alguns pressupostos para o RDS e, desses pressupostos, provém a primeira forma de se estimar prevalências considerando o efeito do desenho, que será apresentado na seção 2.4.1.

A primeira questão relacionada à adequação do modelo markoviano à amostragem RDS foi levantada por Heckathorn (2002), em termos do número ideal de sementes e de convites entregues a cada participante. Ao utilizar uma semente e um convite, o crescimento da amostra seria linear, o que geraria cadeias muitos longas, ou seja, seriam necessárias muitas ondas para obter o tamanho mínimo de amostra calculado. No entanto, a utilização de mais de um convite por pessoa gera uma estrutura de árvore que não corresponde ao modelo linear de cadeias de Markov. Para solucionar esse problema, tem-se que a estrutura de árvore pode ser analisada como uma série de estruturas lineares, ou seja, é possível olhar para cada ramificação dessa árvore como uma cadeia linear. Dessa maneira, se uma análise é válida para uma cadeia linear, então é válida para um conjunto de cadeias lineares.

Ainda com relação ao tamanho das cadeias, dois aspectos são levantados. Por um lado, dependendo do número de pessoas em cada onda, cadeias longas não são necessárias, pois como as probabilidades de transição dependem do tamanho da amostra e não do número de estágios, é possível que o equilíbrio citado anteriormente seja atingido em poucas ondas. Por outro lado, um processo que comece com muitas sementes e tenha apenas uma onda, pode atingir o tamanho de amostra necessário, mas acessa apenas as pessoas mais visíveis na população. Há ainda que se considerar que pessoas intermediárias na amostra (aquelas que não são nem sementes, nem fazem parte da última onda) são recrutadas e recrutam, o que aumenta a heterogeneidade de suas características. Dessa forma, a utilização de um número razoável de ondas, torna todas as pessoas da população acessíveis ao estudo e gera uma amostra equilibrada e heterogênea (Heckahtorn, 2002).

2.4.1. Pressupostos do RDS e a estimação de prevalência com base no equilíbrio da amostra.

O primeiro pressuposto para a boa estimação de prevalências usando RDS é o de que a amostra alcance o equilíbrio. Como já foi citado, o equilíbrio de um processo markoviano é definido como o estado para o qual a matriz de transição de probabilidades converge e torna-se estável, ou seja, sem se modificar de uma onda para a próxima. Um suporte para a noção de que poucas ondas são suficientes para atingir o equilíbrio é uma propriedade matemática de processos markovianos, que diz que esses processos convergem para o equilíbrio a uma taxa geométrica, isto é, muito rapidamente. Dessa forma, as características da amostra não terão dependência com relação às características das sementes, que são escolhidas de forma não aleatória.

A partir desse pressuposto, um dos pontos fracos da utilização do RDS, que é a potencial semelhança com as sementes, fica anulado. Por outro lado, esse equilíbrio traz à mente uma pergunta: quantas ondas são necessárias para que a amostra entre em equilíbrio? Para responder essa pergunta, Heckathorn (1997) recorreu à teoria conhecida como *small world*, segundo a qual, por exemplo, Killworth & Bernard (1978/79) mostram que todas as pessoas de uma população encontram-se indiretamente conectadas numa distância de aproximadamente seis graus de separação. Para testar isso, algumas simulações foram feitas e Heckathorn (1997) constatou que à medida que as ondas foram sendo formadas, a prevalência estimada na amostra foi se estabilizando, sendo que, nesses exemplos, a partir da quinta ou sexta onda, as estimativas não apresentaram modificações significativas. Assim, ele concluiu que era razoável considerar seis ondas como um bom número para a amostra entrar em equilíbrio. Posteriormente, Watts (2003 e 2004) desenvolveu estudos nesse sentido.

Um outro aspecto das redes sociais que também foi considerado por Heckathorn (1997), foi o fato de que algumas relações são mais prováveis do que outras. Ele modela o processo de escolha de um novo participante (filho) como um processo que pode ser gerado por dois critérios de decisão. O primeiro é por *imbreeding* (endogamia). Quando a pessoa usa o critério de endogamia, ela está convidando alguém porque é igual a ela. Se isso acontece, uma pessoa com característica E1 certamente convidará outra pessoa com característica E1. Quando isso não acontece, ela não utiliza este critério de escolha, isso significa que não haverá relação entre a característica dela e a do escolhido, que

será selecionado de forma aleatória dentre todos os seus conhecidos, independendo do grupo ao qual essa pessoa pertença (inclusive, o próprio grupo E1).

Essa noção de *imbreeding* é importante, pois permite a derivação do outro pressuposto do modelo RDS, que diz que para as amostras serem não viesadas é preciso que o *imbreeding* de todos os grupos, ou seja, de todos os estados, seja igual. Ainda assim, há uma outra limitação, que se refere ao fato de que quanto maior for o *imbreeding*, mais longas as cadeias devem ser para que o equilíbrio seja atingindo, ou seja, mais ondas são necessárias. Portanto, a amostra apresenta boas estimativas de prevalência para *imbreeding* controlados. Quando isso não acontece e esses coeficientes são muito elevados, há um indício de que os grupos são bastante isolados. Dessa forma, é aconselhável planejar mais de uma amostra RDS, realizando uma para cada subgrupo identificado.

A primeira maneira de obter estimativas de prevalência para dados gerados a partir de RDS levava em consideração todos esses pressupostos apresentados. Assim:

$$E1 = \frac{P1(1-I2)}{1-(I1+P1(1-I1))+P1(1-I2)} \qquad (5)$$

onde E1 é a estimativa da prevalência, P1 e P2 são as proporções amostrais de pessoas com as características E1 e E2, respectivamente, e I1 e I2 são os parâmetros que medem o *imbreending* em cada estado.

No entanto as suposições necessárias para a utilização da equação 5 são muito fortes (conservadoras). Para relaxar essas suposições, e obter estimativas considerando situações mais factíveis, outro modelo de ponderação foi obtido. Esse modelo será apresentado na seção 2.4.2. Vale destacar que as estimativas obtidas nessa dissertação foram calculadas com base nesse modelo subseqüente.

2.4.2. Introdução ao modelo proposto por Heckathorn (2002), assumindo reciprocidade.

Após a formulação inicial de 1997, alguns pressupostos deste primeiro modelo foram revistos e algumas propostas novas, formalizadas. Heckathorn (2002) analisou uma importante fonte de viés que ocorre em dados obtidos com cadeias de referência, e que provém do *imbreeding*, agora analisado sob o conceito de homofilia. Homofilia é uma medida utilizada para quantificar o quanto os pares recrutados são semelhantes.

Nos casos extremos, a homofilia pode ser perfeita, quando todas as ligações são estabelecidas entre pessoas do mesmo 'tipo', ou nula, quando todas as ligações são estabelecidas entre pessoas com características distintas. A homofilia pode ainda ser positiva ou negativa. Homofilia positiva é quando a proporção de ligações entre pessoas do mesmo tipo é maior do que a proporção de ligações entre pessoas de estados diferentes. Diante da homofilia positiva, demonstra-se que as estimativas podem se tornar superestimadas, como já foi evidenciado no primeiro modelo. O pressuposto de que a amostra RDS é não enviesada apenas quando a homofilia é a mesma para todos os grupos continua sendo necessário, assim como os cálculos para obtenção das estimativas.

Um problema de considerar a modelagem com base na homofilia é a impossibilidade de calculá-la, pois o tamanho da população é desconhecido. Outro problema é a ausência de uma estratégia de controle para os casos onde a homofilia não é a mesma nos diferentes grupos. Com isso, abandona-se o modelo que considera o equilíbrio e passa-se a trabalhar com um modelo baseado nas estimativas das redes de relação das pessoas (Heckathorn, 2002).

Nesse novo modelo proposto, a amostra é utilizada para obter estimativas sobre a forma como as pessoas estão conectadas na população. A partir disso, as proporções populacionais podem ser obtidas. A estimação é feita em duas etapas e, para isso, é preciso compreender o conceito de reciprocidade. Esse conceito e a derivação do modelo de estimação com base neste conceito são apresentados a seguir, sob a forma de um exemplo extraído de Salganik e Heckathorn (2004).

Considere-se uma população hipotética onde as pessoas se dividem em dois grupos, por exemplo, com relação ao status sorológico para o HIV (HIV positivos e HIV negativos). A população é composta de 10 pessoas, 6 do grupo A e 4 do grupo B. Na população há 6 ligações estabelecidas entre pessoas do grupo A com pessoas do grupo B. O esquema abaixo apresenta a estrutura dessas ligações.

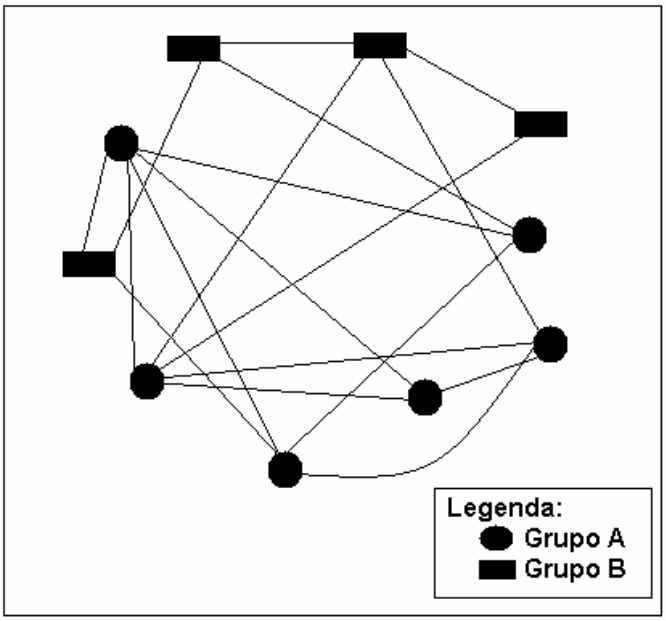

Figura 2.4. Representação de uma população como uma rede de pessoas conectadas, pertencentes a dois grupos, A e B.

A informação sobre a estrutura de ligação entre essas pessoas pode ser armazenada numa matriz X, onde xij é uma variável dicotômica que assume valor 1 se o indivíduo i conhece o indivíduo j, e 0 caso contrário. Em Salganik & Heckathorn (2004), apenas relações recíprocas foram consideradas, ou seja, se A conhece B, B, necessariamente, conhece A. E, assim, tem-se uma matriz X simétrica. Como já foi mencionado também, define-se ainda o grau de uma pessoa (di) como o número de pessoas que ela conhece, tal que $d_i = \Sigma_j x_{ij}$. A matriz X e os graus são apresentados abaixo, considerando como primeiro indivíduo da população o retângulo superior esquerdo e os demais no sentido horário.

$$X = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \end{array} \\ \begin{bmatrix} - & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & - & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & - & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & - & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & - & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & - & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & - & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & - & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & - & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & - \end{bmatrix} \end{array} \quad \text{e} \quad d_i = \begin{bmatrix} 3 \\ 4 \\ 2 \\ 3 \\ 4 \\ 3 \\ 4 \\ 5 \\ 3 \\ 5 \end{bmatrix}$$

O total de ligações das pessoas do grupo A (R_A) é definido pela soma de todas as ligações das pessoas do grupo A, que pode ser dada também pela multiplicação do número de pessoas pertencentes ao grupo A (N_A) pela média dos graus dessas pessoas (D_A). Ou seja: $R_A = \Sigma_{i \in A}\, di = N_A * D_A$.

Como as relações são recíprocas, é possível calcular a probabilidade de uma pessoa do grupo A possuir uma ligação com uma pessoa do grupo B ($C_{A,B}$), tal que:

$C_{A,B} = T_{A,B} / R_A$ e $C_{B,A} = T_{A,B} / R_B$.

Assim, $C_{B,A} * R_B = T_{A,B}$ e $C_{A,B} * R_A = T_{A,B}$ $\therefore$ (6)

$C_{B,A} * R_B = C_{A,B} * R_A$ $\therefore$ $C_{B,A} * N_B * D_B = C_{A,B} * N_A * D_A$.

No entanto, mesmo com a informação completa sobre a rede de conexões das pessoas, ainda é necessário estimar as proporções de pessoas em cada grupo. Dividindo ambas as equações por N tem-se a proporção de pessoas em cada categoria:

$C_{B,A} * PP_B * D_B = C_{A,B} * PP_A * D_A$ onde $PP_A = N_A/N$ e $PP_B = N_B/N$. (7)

Como $PP_A + PP_B = 1$,

$PP_A * (D_A * C_{A,B}) = (1 - PP_A) * (D_B * C_{B,A})$ $\therefore$

$PP_A = D_B * C_{B,A} / (D_A * C_{A,B} + D_B * C_{B,A})$ e (8)

$PP_B = D_A * C_{A.B} / (D_A * C_{A,B} + D_B * C_{B,A})$

Aplicando estas fórmulas à população exemplificada na Figura XY, obtém-se $R_A = 24$ e $R_B = 12$. Nesta população, há 6 ligações entre A e B, isto é, $T_{A,B} = 6$, logo $C_{A.B} = 6/24 = 0{,}25$ e $C_{B,A} = 6/12 = 0{,}5$.

Calculando PPA e PPB, chega-se às estimativas corretas de 0,6 e 0,4, respectivamente. Deve-se lembrar que, neste exemplo, a estimativa amostral foi igual ao valor populacional, pois a população é completamente conhecida.

O exemplo mostra então que é possível estimar a proporção nos grupos A e B, mas é necessário dispor de informações sobre as redes de conexão das pessoas. O próximo passo mostra como derivar essas estimativas a partir de dados da amostra.

2.4.3. Estimação de prevalências a partir do modelo de reciprocidade

Considere a amostragem como um processo de escolha de vértices $NI(j)_{w=x}$ e ligações $EI(e_{j\to k})_{r=x}$. Os vértices são as pessoas e as ligações, são geradas pelo processo de recrutamento. O vértice 0 é a semente, que se conecta ao vértice 1 (pessoa da primeira onda) através de uma ligação. Duas funções indicadoras podem ser definidas para a representação dos vértices e ligações:

$$NI(j)_{w=x} = \begin{cases} 1, \text{se o vérice j é selecionado na onda x} \\ 0, \text{caso contrário} \end{cases}$$

$$EI(e_{j\to k})_{r=x} = \begin{cases} 1, \text{se a ligação entre j e k é feita durante o recrutamento x} \\ 0, \text{caso contrário} \end{cases}$$

Assume-se que todas as pessoas formam uma única rede, ou seja, qualquer par de pessoas está conectado, mesmo que de forma indireta, ou seja, mesmo que seja necessário percorrer um caminho entre algumas pessoas até a conexão ocorrer de fato. Formalmente, a rede de interesse é um grafo com um único componente.

Assume-se também que cada pessoa que recebe 1 (um) convite, utiliza-o de forma correta; e considera que a pessoa a ser recrutada é escolhida aleatoriamente entre as pessoas conhecidas, ou seja, sem considerar a homofilia, por exemplo. Assim, a probabilidade de um vértice j, selecionado na onda x, estabelecer ligação com outro vértice k, para formar a onda x+1 é dada por:

$$P[EI(e_{j\to k})_{r=x+1} = 1 \mid NI(j)_{w=x} = 1] = \frac{1}{d_j} \qquad (9)$$

Além disso, assume-se que as sementes são escolhidas de forma proporcional aos seus graus, por serem escolhidas pelos pesquisadores (que não são necessariamente membros da população alvo). Assim, considerando todos os membros da população, uma pessoa com 20 amigos tem o dobro de chance de ser escolhida como semente do que uma pessoa que tenha apenas 10 amigos. Matematicamente,

$$P[NI(j)_{w=0} = 1] = \frac{d_j}{\sum_{i\in N} d_i} \qquad (10)$$

Algumas consequências dessas considerações precisam ser apresentadas. Uma vez que se determinou a probabilidade de seleção de uma semente, e considerando a

informação sobre seu grau, é possível determinar a probabilidade de realizar a primeira ligação entre j e k, tal que:

$$P[EI(e_{j\to k})_{r=1}=1]=P[EI(e_{j\to k})_{r=1}=1\,|\,NI(j)_{w=0}=1]\times P[NI(j)_{w=0}=1]\therefore$$

$$P[EI(e_{j\to k})_{r=1}=1]=\frac{1}{d_j}\times\frac{d_j}{\sum_{i\in N}d_i}=\frac{1}{\sum_{i\in N}d_i} \quad (11)$$

Ou seja, se os primeiros vértices (sementes) são escolhidos com probabilidade proporcional ao grau, então a probabilidade de cada ligação j→k será a mesma no primeiro recrutamento. O próximo passo é então derivar a probabilidade de selecionar os vértices no primeiro recrutamento.

$$P[NI(j)_{w=1}=1]=\sum_{d_j}\frac{1}{\sum_{i\in N}d_i}=\frac{d_j}{\sum_{i\in N}d_i} \quad (12)$$

A equação expressa que a probabilidade de seleção do novo vértice é proporcional ao grau da pessoa j. Com isso,

$$P[EI(e_{j\to k})_{r=1}]=P[EI(e_{j\to k})_{r=2}]=\frac{1}{\sum_{i\in N}d_i} \quad (13)$$

As equações acima assumem que as probabilidades de seleção de um vértice e de uma ligação se mantêm constantes, independente da onda em que se encontra o processo de amostragem. Com essa informação, é possível começar a derivar as estimativas amostrais para as proporções em cada grupo.

A informação sobre o grau de cada pessoa é obtida durante o processo de amostragem, quando cada participante responde perguntas relacionadas ao tamanho de sua rede social, e é de posse dessa informação que começa o procedimento de estimação.

Já foi mostrado que a proporção populacional de um grupo com característica A é dada por: $PP_A = D_B{*}C_{B,A} / (D_A{*}C_{A,B} + D_B{*}C_{B,A})$

Assim, o primeiro passo é estimar as probabilidades cruzadas, ou seja, $C_{A,B}$ e $C_{B,A.}$ Esse cálculo é feito com base nas relações estabelecidas entre recrutador e recrutado. Uma vez que a amostra tenha sido obtida, é possível calcular r_{AA}, r_{AB}, r_{BA} e r_{BB}, que são as estimativas do número de ligações feitas ente pessoas do grupo A com

pessoas do grupo A, pessoas do grupo A com pessoas do grupo B e assim por diante. Com isso:

$$\hat{C}_{A,B} = \frac{r_{AB}}{r_{AB} + r_{AA}} \quad \text{e} \quad \hat{C}_{B,A} = \frac{r_{BA}}{r_{BA} + r_{BB}} \tag{14}$$

O próximo passo é estimar D_A e D_B, ou seja, a média dos graus das pessoas do grupo A e do grupo B. A média aritmética não é um bom estimador nos casos de cadeias de referência, pois pessoas com altos graus tendem a ficar super-representadas.

Assim, uma forma de construir um estimador assintoticamente não enviesado para D_A é utilizar o processo de estimação de Hansen-Hurwitz, que será abordado a seguir e está apresentado em Salganik & Heckathorn (2004). Um estimador não enviesado é aquele cuja estimativa é o verdadeiro valor do parâmetro, ou seja, o valor populacional. Já um estimador assintoticamente não enviesado é aquele que converge para o valor populacional conforme o n aumenta, ou seja, quanto maior o tamanho da amostra, melhor a estimativa.

O processo de Hansen-Hurwitz consiste em atribuir pesos para cada elemento na amostra utilizando o inverso da probabilidade de ser sorteado. Com ele:

$$\hat{D}_A = \frac{\hat{R}_A}{\hat{N}_A} = \frac{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{1}{p_i} \times d_i}{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{1}{p_i}} \tag{15}$$

onde p_i é a probabilidade de uma pessoa i ser selecionada em determinado recrutamento. Essa probabilidade é desconhecida. Porém, como as pessoas são escolhidas de forma proporcional ao grau, a probabilidade relativa de escolha para dois vértices, j e k é dada por:

$$p_i = \frac{d_i}{\sum_{j \in N} d_j} \quad \text{e} \quad p_k = \frac{d_k}{\sum_{j \in N} d_j} \tag{16}$$

$$\frac{p_k}{p_i} = \frac{d_k}{\sum_{j \in N} d_j} \times \frac{\sum_{j \in N} d_j}{d_i} = \frac{d_k}{d_i} \Rightarrow p_i = \frac{d_i p_k}{d_k} \tag{17}$$

Substituindo na equação anterior, tem-se o estimador para a média dos graus no grupo A:

$$\hat{D}_A = \frac{\hat{R}_A}{\hat{N}_A} = \frac{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{d_k}{d_i p_k}\times d_i}{\frac{1}{n_A}\sum_{i=1}^{n_A}\frac{d_k}{d_i p_k}} = \frac{\frac{d_k}{d_i p_k}\sum_{i=1}^{n_A}\frac{d_i}{d_i}}{\frac{d_k}{d_i p_k}\sum_{i=1}^{n_A}\frac{1}{d_i}} = \frac{\sum_{i=1}^{n_A}1}{\sum_{i=1}^{n_A}\frac{1}{d_i}} \quad (18)$$

$$\hat{D}_A = \frac{n_A}{\sum_{i=1}^{n_A}\frac{1}{d_i}} \quad (19)$$

E da mesma forma, o estimador para a média dos graus no grupo B é $\hat{D}_B = \frac{n_B}{\sum_{i=1}^{n_B}\frac{1}{d_i}}$.

Combinando as equações dessa seção, já é possível calcular as estimativas de prevalência para dois grupos utilizando o modelo de reciprocidade, que serão dadas por:

$$PP_A = \frac{\hat{D}_B \times \hat{C}_{B,A}}{\hat{D}_A \times \hat{C}_{A,B} + \hat{D}_B \times \hat{C}_{B,A}} \text{ e } PP_B = \frac{\hat{D}_A \times \hat{C}_{A,B}}{\hat{D}_A \times \hat{C}_{A,B} + \hat{D}_B \times \hat{C}_{B,A}} \quad (20)$$

É importante lembrar que novos métodos têm sido propostos para obtenção de estimativas de prevalência (Volz & Heckathorn, 2008; Heckathorn, 2007). No entanto, a escolha desse método para compor essa dissertação se baseia no fato de que os trabalhos de Heckathorn (1997 e 2002) e Salganik & Heckathorn (2004) constituem a base teórica fundamental do RDS e ainda têm sido bastante considerados nos processos de estimação.

2.5. Métodos de simulação de amostragem RDS.

Uma simulação computacional é um método implementado em um computador com o objetivo de reproduzir um processo real e explorar algumas de suas propriedades, que dificilmente seriam observáveis empiricamente (Hartmann, 2005). Na literatura são apresentadas algumas poucas estratégias para a geração de populações simuladas organizadas em redes, para avaliação das propriedades estatísticas do RDS. Isso acontece, principalmente, por ser difícil realizar estudos analíticos que considerem todo o conjunto de pressupostos da metodologia, ainda mais no que se refere diretamente à questão da amostragem sem reposição (Gile & Handcock, 2009). Duas abordagens serão apresentadas a seguir. No entanto, deve-se considerar que, em ambas, os exemplos apresentados foram conduzidos utilizando amostragem com reposição.

Em uma dessas vertentes, as redes simuladas são geradas utilizando dados de dados reais gerados pela implementação de RDS, e esse processo é conhecido como *data-driven*. Um exemplo desse tipo de simulação é apresentado por Salganik (2006) para avaliar possíveis vieses produzidos por amostras RDS, utilizando uma proposta de *bootstrap* modificado. O trabalho propôs o seguinte algoritmo (aqui exemplificado com um exemplo de estudo desenvolvido anteriormente no Brasil):

- Parte-se de uma amostra de dados empíricos, gerada por RDS, por exemplo, 500 HSH que vivem na região metropolitana de Campinas, na qual é de interesse estimar uma proporção, por exemplo, prevalência de HIV positivos. As pessoas da amostra possuem um atributo dicotômico, A e B, que influencia a forma com que o processo de recrutamento ocorre (homofilia).

- Para realizar a simulação, dois conjuntos são definidos. Cada elemento da amostra é observado e classificado de acordo com o atributo do recrutador, originando-se os conjuntos de pessoas convidadas por pessoas do tipo A, A_{rec} e convidadas por pessoas do tipo B, B_{rec}.

- O processo de reamostragem é iniciado. Para isso, **s** pessoas são selecionadas aleatoriamente para servirem de sementes e um número **c** de convites é escolhido para passar a cada participante.

- Para cada semente selecionada, é verificado o grupo ao qual ela pertence (A ou B). Se a semente for do tipo A, as **c** pessoas são selecionadas aleatoriamente do conjunto A_{rec}, e se ela for do tipo B, as **c** pessoas são amostradas do conjunto $B_{rec.}$ O total de pessoas selecionadas por todas as sementes compõe a primeira onda.

O procedimento é repetido até que uma amostra com o mesmo tamanho da amostra original seja obtida. Vale destacar que a amostra final foi obtida com reposição, o que não ocorre na prática. Além disso, é preciso considerar também que a qualidade desse processo dependerá também da qualidade dos dados originais obtidos, ou seja, se os dados não forem heterogêneos e não representarem bem a população, o processo de reamostragem também não representará.

Uma outra vertente para a geração de dados simulados se baseia em métodos para a geração das redes a partir de regras de formação. Esses métodos, conhecidos como *model-driven*, são baseados em modelos matemáticos, ou seja, as redes são

geradas a partir de regras que definem a probabilidade de dois dos seus vértices se conectarem, ou a probabilidade de haver ligação entre duas pessoas. Salganik & Heckathorn (2004) propõem o seguinte algoritmo para avaliar estimativas de prevalência obtidas em amostras geradas por RDS:

- Define-se os Parâmetros do modelo gerador:

  - Número de pessoas na população, n, e proporção de pessoas no grupo A e B;

  - Distribuição dos graus (tamanho da rede pessoal de conhecidos) da superpopulação; A partir dessa distribuição que os vértices são amostrados ;

  - Nível de interconectividade, I, entre os dois grupos, A e B, definido como a razão entre o total de ligações cruzadas ($T_{A,B}$) e o mínimo entre o total de ligações de cada grupo (R_A e R_B).

- Para cada pessoa i na população n, é gerado um número aleatório com base na distribuição de graus da superpopulação. Esse número (d_i) representa o número de ligações dessa pessoa, ou seja, seu grau.

- Para cada pessoa i define-se o nível de interconectividade (I) e o número de ligações cruzadas ($T_{A,B}$) é calculado.

- Em seguida, são simuladas as ligações entre as pessoas. Primeiro são feitas as ligações cruzadas, de forma aleatória entre “indivíduos” do grupo A e “indivíduos” do grupo B. Depois, verifica-se o número de ligações restantes para cada “indivíduo” e são realizadas ligações dentro do grupo.

- Algumas vezes, esse algoritmo pode não fechar, ou seja, podem sobrar ligações sem que haja pessoas para conectar. Nesses casos, o processo deve ser reiniciado. De forma geral, o processo de simulação da população está completo com os passos anteriores e amostra pode ser gerada utilizando RDS. Para isso, basta escolher um número s de sementes, que iniciarão a amostra e um número c de convites entregues a cada participante.

As populações e amostras simuladas nessa dissertação foram geradas segundo um algoritmo *model-driven*, como será apresentado no capítulo 4, e se assemelham ao modelo acima mencionado. A vantagem de se utilizar essa abordagem, em relação à

abordagem *data-driven* é que aqui é possível criar as populações de diferentes formas e estabelecer os parâmetros desejados, o que não é possível quando a base da simulação parte de dados gerados a partir de um estudo anterior.

3. Objetivos

3.1. Objetivo Geral

Avaliar, através de simulações, a precisão de estimativas de prevalência de doenças transmissíveis, obtidas utilizando a metodologia de amostragem "Respondent-Driven Sampling" e o modelo de estimação proposto por Heckathorn (2002), em populações organizadas em redes complexas.

3.2. Objetivos Específicos

- Construir algoritmo de geração de diferentes cenários de populações organizadas em redes complexas, através de técnicas de simulação computacional e utilizando como referência dados empíricos.
- Implementar, na plataforma R, o modelo de estimação de prevalência em amostras RDS, proposto por Heckathorn (2002).
- Simular diferentes implementações da metodologia de amostragem RDS e analisar as estimativas obtidas em função das características das redes subjacentes e da própria forma de implementação do RDS.

4. Metodologia

O primeiro passo que será descrito se refere ao método de geração da população a ser amostrada, a qual está organizada sob a forma de uma rede de conhecidos. A escolha dos parâmetros utilizados em uma simulação é muito importante, pois quanto mais esses parâmetros estiverem próximos das características reais de uma população, o mesmo acontecerá com os resultados obtidos através dessas simulações. Assim, os parâmetros utilizados para a geração da população foram escolhidos a partir de dados reais obtidos pelo Projeto "Semear Saúde", descrito na seção 4.1. Esse projeto teve como população-alvo, os homens que fazem sexo com homens (HSH), residentes em Campinas, SP, nos anos de 2005 e 2006, e seguiu corretamente as orientações da metodologia RDS (Mello *et al*., 2008). É importante destacar que a presente dissertação não tem o objetivo de apresentar ou discutir os resultados do Projeto "Semear Saúde", mas apenas identificar e utilizar as variáveis associadas à estrutura de rede desta população, de modo a parametrizar as simulações.

Em seguida, nas seções 4.2. e 4.3., são descritos: o algoritmo de simulação das populações organizadas em redes; o algoritmo de distribuição da característica a ser estimada, isto é, o status infectado ou não infectado de cada indivíduo; e o processo de obtenção das amostras por RDS para geração das estimativas de prevalência da infecção, com a implementação do modelo descrito em Heckathorn (2002). A dissertação foi toda desenvolvida em R, versão 2.7.1 (R, 2008), e os *scripts* utilizados estão referenciados e apresentados nos anexos.

De forma geral, o fluxograma de desenvolvimento das simulações apresentadas nessa dissertação está apresentado a seguir, na figura 4.1.

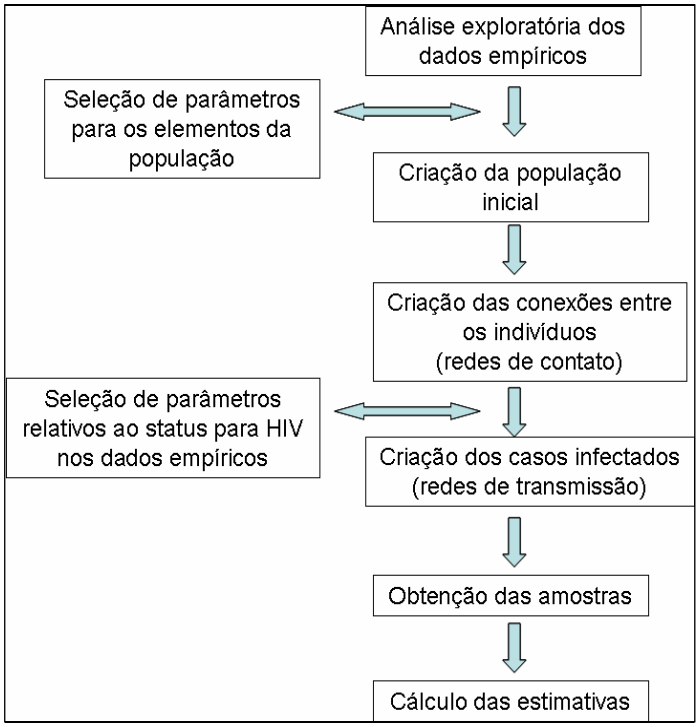

Figura 4.1. Algoritmo utilizado para as simulações.

4.1. Análise exploratória dos dados empíricos

4.1.1. Descrição do Projeto "Semear Saúde"

O Projeto "Semear Saúde" foi um estudo com desenho seccional, cujo público-alvo era a população HSH residente na região metropolitana de Campinas, SP (Mello *et al.*, 2008). Essa região é composta pelos seguintes municípios: Artur Nogueira, Engenheiro Coelho, Cosmópolis, Holambra, Santo Antônio da posse, Americana, Paulínia, Jaguariúna, Pedreira, Santa Bárbara do Oeste, Nova Odessa, Sumaré, Hortolândia, Monte Mor, Indaiatuba, Valinhos, Vinhedo, Itatiba e Campinas. Os critérios de inclusão estabeleciam que os participantes deveriam ser maiores de 14 anos e ter praticado sexo oral ou anal com outro homem nos últimos seis meses.

A amostra foi obtida com a utilização da técnica RDS e incluiu uma entrevista auto-respondida, utilizando o método ACASI (Audio Computer Assisted Self Interview) (Simões *et al.*, 2006), teste rápido para sífilis (obrigatório), teste rápido para HIV (opcional) e aconselhamentos pré e pós-teste. O local escolhido para sediar o estudo ficava em Campinas e era de fácil acesso, de acordo com o que é desejável para a utilização do RDS. Os participantes precisavam comparecer ao local do estudo apenas duas vezes. A primeira quando participavam e, posteriormente, depois de

recrutarem outras pessoas, para o recebimento do incentivo secundário. Os dados foram coletados entre 25 de outubro de 2005 e 21 de outubro de 2006.

O estudo foi realizado pelo Instituto Horizons/Population Council, em parceria com diversos órgãos. Alguns aspectos éticos devem ser citados. Seu protocolo foi aprovado pelo *Institutional Review Board of the Population Council*, nos Estados Unidos, pelo Comitê de Ética da Universidade de Campinas (UNICAMP) e pelo Conselho Nacional de Ética em Pesquisa (CONEP). Todos os participantes assinaram o termo de consentimento livre e esclarecido, e para aqueles entre 14 e 18 anos, era necessário também o consentimento de um responsável legal, com exceção aos casos em que revelar a condição de HSH a esse responsável poderia gerar possíveis estigmatizações ou retaliações por parte do mesmo.

Alguns dos objetivos desse estudo eram estimar a soro prevalência da infecção pelo HIV e sífilis, além de conhecer o perfil dessa população. A amostra planejada deveria ser composta por 1800 participantes. Ao final do período, 689 HSH haviam sido recrutados, dos quais 658 eram elegíveis e efetivamente participaram do estudo.

O estudo contou com trinta sementes, o que possibilitou que as ondas de recrutamento se estendessem até a vigésima-quarta onda. As oito primeiras ondas contaram com mais de trinta pessoas em cada uma delas, representando 58,15% da amostra. Por outro lado, a partir da décima nona onda, todas tinham menos de dez participantes.

A base de dados desse estudo é bastante diversificada, com informações sobre características sócio-demográficas, identidade e orientação sexual, auto-estima, visibilidade na população-alvo, exposição a atividades de prevenção em HIV, comportamentos sexuais, etc. A partir da análise da distribuição amostral dessas variáveis foram escolhidos os parâmetros utilizados nas simulações.

4.1.2. Construção da base de dados pareados.

Para se trabalhar com amostras que utilizam cadeias de referência, é necessário ter em mãos um banco de dados com informações pareadas, ou seja, um banco de dados onde cada registro (linha) não se refira exatamente a uma pessoa, mas sim a um par de pessoas (recrutador - recrutado). Nesse sentido, em vez de utilizar a base de dados original, que continha informações individuais em cada registro, uma nova base foi elaborada. Nessa base, apenas as variáveis de interesse foram selecionadas, e cada registro dispôs de informações sobre o participante e sobre o seu

recrutador. Com isso, tornou-se possível analisar, por exemplo, como se dá a relação entre as idades dos participantes e seus recrutadores. A tabela 4.1 apresenta as variáveis que compuseram essa base de dados.

Tabela 4.1. Variáveis pertencentes à base de dados pareados.

Variáveis referentes ao participante	Variáveis referentes ao recrutador
Onda	
ID RDS	
Religião	Religião
Estado civil	Estado civil
Cidade de residência	Cidade de residência
Escolaridade	Escolaridade
# de conhecidos na pop. alvo	# de conhecidos na pop. alvo
# de pessoas que convidaria	# de pessoas que convidaria
Status sorológico para HIV	Status sorológico para HIV
Status sorológico para Sífilis	Status sorológico para Sífilis
Orientação sexual	Orientação sexual
Idade	Idade
Raça	Raça
Com quem mora	Com quem mora
Tipo de moradia	Tipo de moradia
Classe econômica	Classe econômica
Renda	Renda

4.1.3. Análise exploratória dos dados.

Foram analisadas as seguintes informações: distribuição dos graus dos participantes para investigação do modelo adequado para sua representação, caracterização das variáveis associadas com homofilia, isto é, características do recrutador que afetam a probabilidade de recrutamento de uma pessoa. Essas características empíricas são relevantes para a parametrização do modelo de simulação de populações em rede.

Distribuição de frequência dos graus individuais (tamanho das redes de contato).

A primeira variável investigada foi o tamanho da rede de conhecidos dos participantes. Para isso utilizou-se a pergunta: “Quantos HSH você conhece, que poderia entrar em contato pessoalmente ou por telefone e que você tenha encontrado no último mês?” do questionário de inclusão do estudo de Campinas. Vale destacar que a mesma investigação foi realizada utilizando uma outra pergunta: “Destes, quantos você convidaria para o estudo?”, mas como os resultados foram semelhantes a primeira foi utilizada para gerar o grau de cada indivíduo da população virtual. O

tamanho médio das redes pessoais foi de 21,98 pessoas, com variação entre 0 e 700 – e os valores extremos foram citados por uma única pessoa cada. A figura 4.2 apresenta a distribuição de freqüências dessa variável.

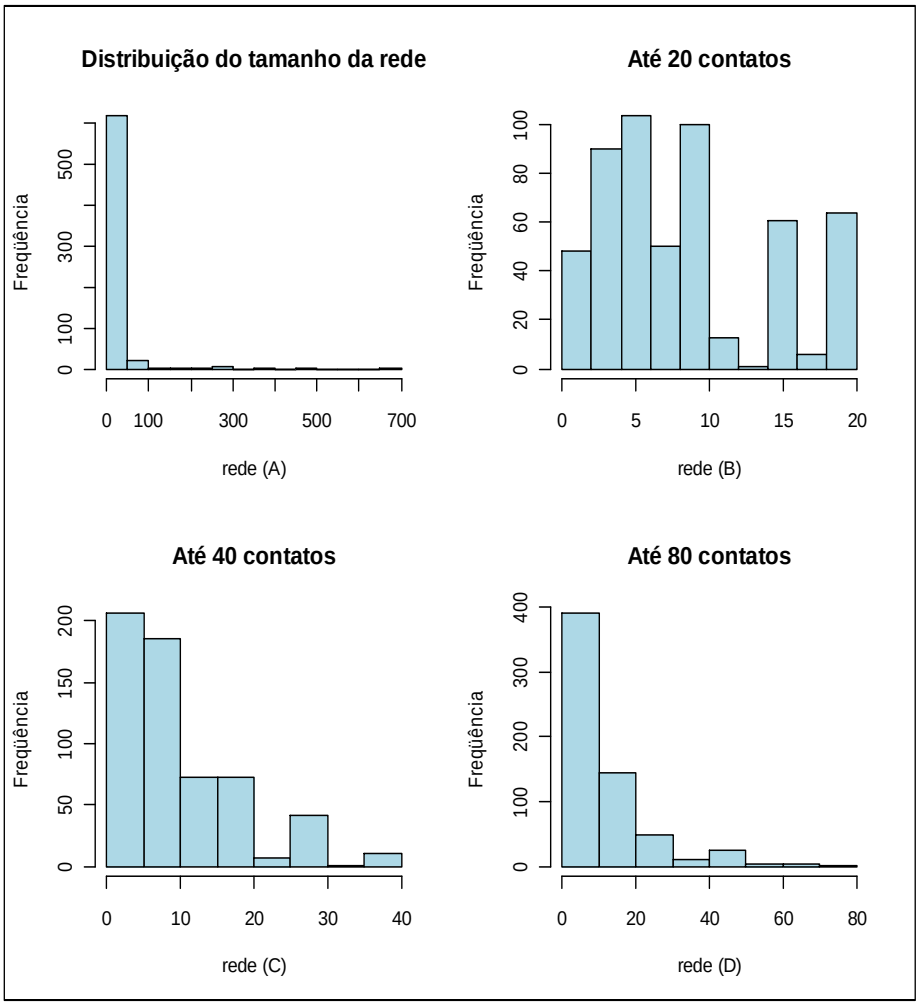

Figura 4.2. Distribuição dos graus (tamanho da rede de conhecidos) dos participantes do Estudo Semear Saúde. (A) Todos os participantes; (B) Restrito àqueles com até 20 conhecidos (80% dos participantes); (C) Restrito àqueles com até 40 conhecidos (90% dos participantes); (D) Restrito àqueles com até 80 conhecidos (95,7% dos participantes).

Pela figura 4.2, é possível concluir que existem muitas pessoas que têm poucos contatos, ou seja, que têm redes de conhecidos pequenas, e poucas pessoas com grandes redes de conhecidos. Assim, o próximo passo foi determinar o melhor modelo para representar a estrutura da distribuição dos graus dos participantes, bem como os seus parâmetros. Como é possível observar na figura 4.2, 95% dos participantes respondeu ter grau de até 80 contatos, e para estabilizar a estimação dos parâmetros do modelo de distribuição de graus, nos restringimos a este subconjunto

da população original para o seu ajuste. Estes dados, reorganizados em intervalos de classe de tamanho 10, são reapresentados na figura 4.3. Essa figura sugere que a distribuição dos graus dos participantes não é linear, e dois modelos alternativos foram investigados, um modelo de lei de potência e um modelo exponencial, ambos apresentados na seção 2.1.

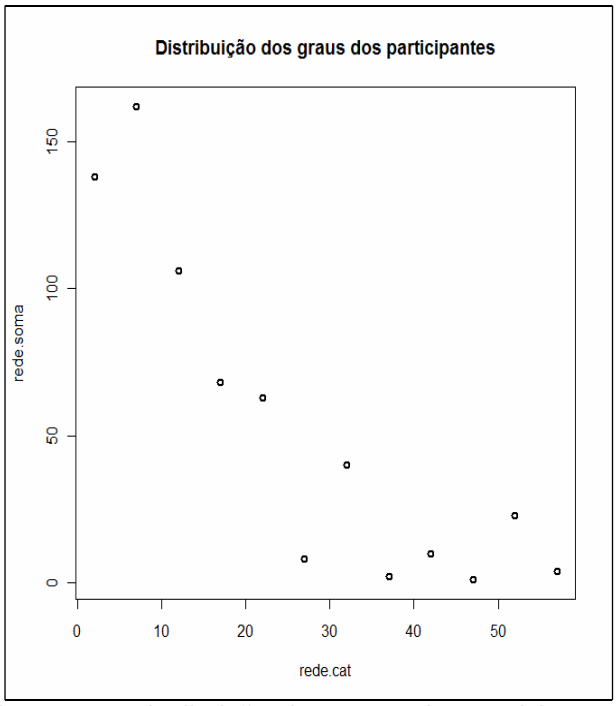

Figura 4.3. Distribuição dos graus dos participantes do estudo empírico.

Para modelar a distribuição de graus, modelos livre de escala ou de potência são utilizados com freqüência, por serem observados com freqüência em sistemas empíricos de redes (Stumpf & Wiuf, 2005, Laird & Jensen, 2006). Uma alternativa também considerada foi o modelo exponencial, que segundo Strogatz (2001), também representa um modelo capaz de produzir bons resultados. A identificação do melhor modelo pode ser feita através do ajuste de uma reta aos dados plotados em um gráfico em escala semi-log (modelo exponencial) ou log-log (modelo livre de escala). Após a análise de diagnóstico dos dois modelos, verificou-se que o modelo exponencial ajustou melhor aos dados avaliados. A tabela 4.2 e a figura 4.4 a seguir, apresentam os resultados desses ajustes.

Tabela 4.2. Ajuste do modelo exponencial e de potência à distribuição dos graus dos participantes do estudo empírico.

Modelo "Exponencial"					**Modelo "Lei de potência"**				
	β	DP(β)	t	p-valor		β	DP(β)	t	p-valor
Intercepto	0.24465	0.63081	8.314	8.39e-06	Intercepto	6.991	1.2085	5.785	0.000177
rede.cat	-0.07517	0.01846	-4.073	0.00224	Log(rede.cat)	-1.2854	0.3749	-3.429	0.006453
R-ajustado	0.5863				R-ajustado	0.4944			
F	0.0022				F	0.0064			

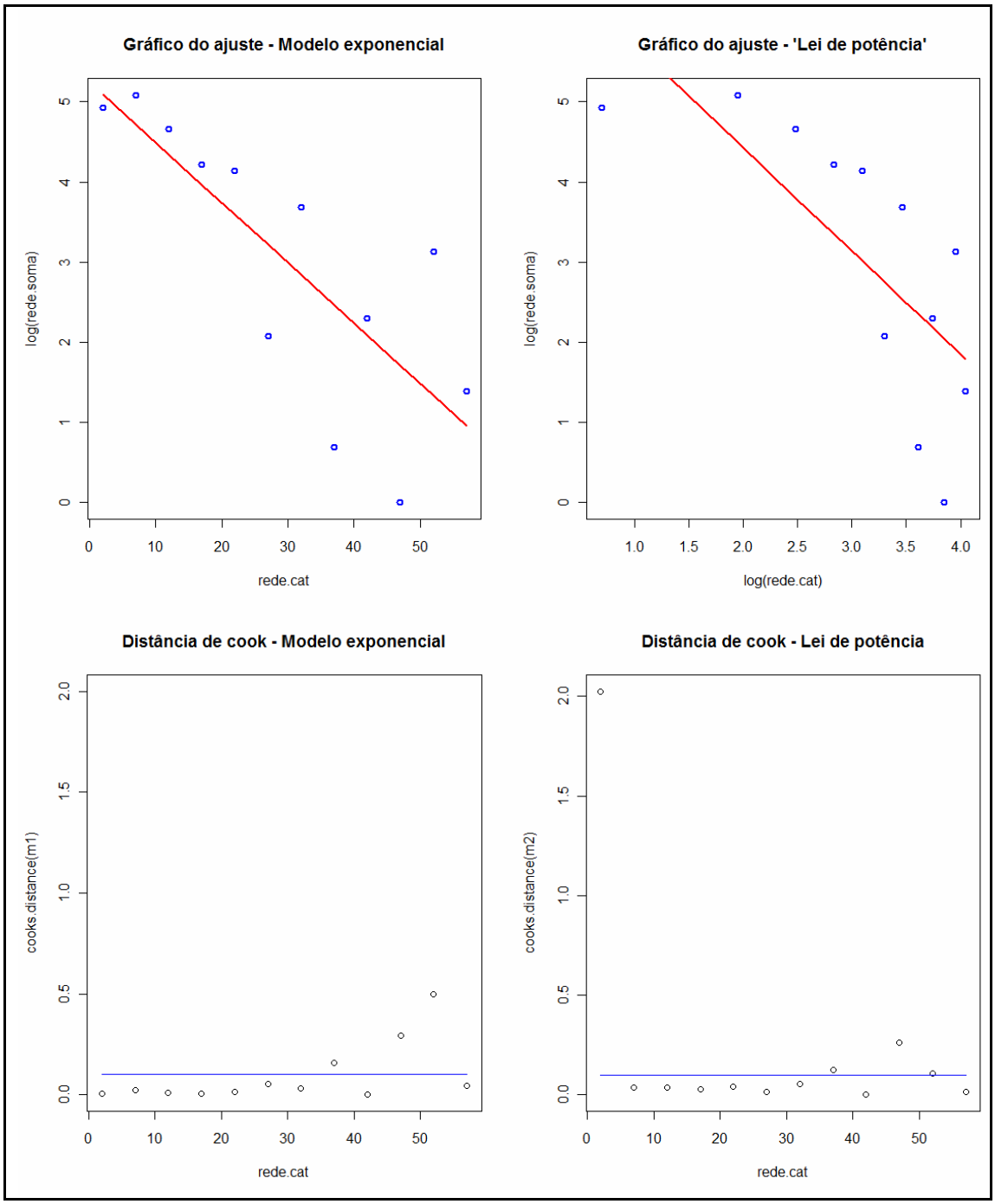

Figura 4.4. Diagnóstico do ajuste do modelo exponencial e do modelo Lei de potência à distribuição dos graus dos participantes do estudo empírico.

Dessa forma, o modelo exponencial com parâmetro 0,08 foi adotado para geração do grau nas populações simuladas, e um número aleatório com a distribuição apresentada na equação 21 foi gerado para atribuir o número de ligações de cada indivíduo:

$$(grau)_i = e^{-(-0,08*j)} = e^{(0,08*j)}. \quad (21)$$

onde j é um valor gerado aleatoriamente.

Características avaliadas para a construção das ligações entre os elementos da população virtual.

Relações sociais em geral são caracterizadas por homofilias, isto é, pessoas tendem a conhecer/contactar pessoas que compartilham características com elas. Para incorporar esta propriedade homofílica nas populações simuladas, o passo subsequente foi a caracterização dos pares "recrutador-recrutado" observada nos dados empíricos, e identificação de características que possivelmente relacionam os participantes e seus recrutadores. Essas informações foram utilizadas posteriormente para gerar as relações de ligação entre os elementos da população virtual.

Para as variáveis contínuas (idade, escolaridade e renda), foi calculada a correlação de Pearson entre as medidas do participante e seu recrutador e em seguida procedeu-se a um teste de correlação. Para as variáveis categóricas (orientação sexual e raça), foi utilizado o teste Qui-quadrado. A tabela 4.3 apresenta os resultados dos testes realizados.

Tabela 4.3. Associação entre atributos do recrutado e do recrutador no estudo empírico.

Variável	Estatística de teste	Graus de liberdade	p-valor
Idade	13,5948	626	< 0,01
Escolaridade	5,7365	623	< 0,01
Orientação sexual	10,08	1	< 0,01
Raça	1,2989	1	0,254
Renda	-0,352	555	0,725

Para as variáveis contínuas que se mostraram significativas ─ idade e escolaridade ─, foram ajustados modelos lineares para estimar a magnitude desta relação. O objetivo era encontrar o parâmetro que seria utilizado posteriormente. O modelo ajustado para a escolaridade não se mostrou significativo e, por isso, essa variável foi excluída do processo de simulação.

Com relação às idades dos participantes e seus recrutadores, o ajuste com os dados empíricos mostrou que, conforme a idade do participante aumenta em um ano, a idade do seu recrutador aumenta em 0,453, ou seja, os recrutadores tendem a convidar participantes um pouco mais velhos (Tabela 4.5).

Dentre as variáveis categóricas, apenas a orientação sexual mostrou-se significativa. Na tabela 4.4, é possível ver a maioria dos participantes do projeto "Semear Saúde" eram homossexuais, que convidaram outros homossexuais. Entre os participantes bissexuais, é possível perceber que aproximadamente 2/3 dos participantes foram também recrutados por homossexuais.

Tabela 4.4. Relação entre a orientação sexual do participante do estudo empírico e seu recrutador

		Participante	
		Homossexual	Bissexual
Recrutador	Homossexual	356	95
	Bissexual	90	48

Tabela 4.5. Ajuste da idade do participante do estudo empírico em relação a idade do seu recrutador

	β	DP(β)	t	p-valor
Intercepto	13.57804	0.91204	14.89	< 0,01
Idade	0.45289	0.03331	13.6	< 0,01
R-ajustado	0.2267			
F	<2e-16			

Dessa forma, os indivíduos que compõem a população simulada receberam atributo "idade" (em anos) e "orientação sexual" (categorizada como homossexual e bissexual), respeitando a distribuição empírica observada para estas variáveis, assim como a distribuição empírica das variáveis dos recrutados condicionada às variáveis dos recrutadores. O procedimento para a inclusão dessas informações nas simulações será apresentado nas próximas seções.

Características associadas com status HIV positivo

Quais são as variáveis que se relacionam com o status sorológico para o HIV nos dados empíricos? E como se dá essa relação? Para responder a essas perguntas, as mesmas variáveis apresentadas acima (tabela 4.3) foram reanalisadas, desta feita, relacionando-as com o status sorológico do participante. Foram utilizados os testes de

Fisher, Qui-quadrado e t de Student, dependendo da situação. Os resultados estão apresentados na tabela 4.6.

Tabela 4.6. Associação entre o status sorológico para HIV do participante do estudo empírico e variáveis sócio-demográficas .

Variável	Estatística (g.l.)	p-valor
Idade	4,407 (57)	< 0,01
Classe econômica	*	< 0,01
Status sorológico para Sífilis	6.2565 (1)	0.0124
Escolaridade	-2.3698 (57)	0.0212
Orientação sexual	*	0.0491
Raça	*	0.6980

* A estatística de teste para o teste exato de Fisher não é apresentada no R.

Para aquelas variáveis que foram estatisticamente significativas (p-valor < 0,05), foi ajustado um modelo de regressão logística que determinou os coeficientes utilizados para o cálculo da probabilidade de ser um elemento infectado na população virtual. O procedimento para o cálculo dessa probabilidade será apresentado nas próximas seções. A Tabela 4.7 mostra os parâmetros do modelo logístico ajustado.

Tabela 4.7. Ajuste do modelo logístico para determinação das variáveis de influência no status sorológico para HIV do participante do estudo empírico.

	OR	β	DP(β)	z	p-valor
Intercepto		-2,28631	0,75779	-3,017	< 0,01
Idade	1,0790	0,076	0,01856	4,094	< 0,01
Classe econômica (2)	0,3026	-1,19519	0,43806	-2,728	< 0,01
Classe econômica (3)	0,2469	-1,39886	0,48089	-2,909	< 0,01
Classe econômica (4)	0,5068	-0,67969	0,7247	-0,938	0,348
Escolaridade	0,8950	-0,11091	0,05541	-2,002	0,045
Deviance: 297,57		df: 554			
AIC: 281,3					

4.2. Algoritmo de Geração das populações virtuais e casos infectados.

As populações simuladas foram geradas segundo um algoritmo *model-driven*, que se assemelha ao segundo modelo apresentado na seção 2.5. As diferenças se referem à determinação da distribuição dos graus das pessoas – isto é, a distribuição do número de conhecidos das pessoas –, da seleção das pessoas que fariam parte de cada grupo (A/ B ou infectados/ não infectados) e da realização da ligação entre as pessoas.

O processo de criação de populações simuladas foi dividido em três etapas: 1) a construção de uma população inicial de indivíduos com atributos, mas sem ligações entre si; 2) a criação das redes sociais a partir da ligação entre elementos da população virtual; 3) atribuição de status “infectado/ não infectado” aos indivíduos da população. Todo o processo foi desenvolvido em R e está apresentado no Anexo I.

Criação de uma população virtual inicial

O primeiro passo foi a criação de uma população virtual inicial composta por 25.000 indivíduos. Essa população foi criada a partir dos dados empíricos, utilizando amostragem aleatória simples com reposição, porém guardando-se apenas as variáveis: idade, classe econômica e escolaridade. A orientação sexual foi atribuída de acordo com a proporção de homossexuais e bissexuais nos dados empíricos de Campinas. O grau de cada elemento da população simulada foi em seguida determinado por amostragem da distribuição exponencial apresentada na equação 21. O algoritmo utilizado para geração desta população pode ser encontrado no Anexo I (A).

4.2.1 Simulação das redes de contato social.

Tendo sido criada a população, o próximo passo consistiu em estabelecer as relações de contato entre os elementos dessa população, isto é, definir “quem conhece quem”. Como será visto até o final dessa seção, ao final, foram geradas 16 populações, compostas por 25.000 elementos cada.

Para avaliar como a estrutura das redes sociais pode afetar as estimativas geradas pelo RDS, quatro processos geradores de ligação entre os elementos das populações foram considerados. Esses processos foram denominados Ligação 01, 02, 03 e 04 e a forma como foram geradas será apresentada a seguir.

Na ligação 01 os elementos foram conectados de forma aleatória, respeitando apenas o grau individual estabelecido previamente. Isto é, não há homofilia. O *script* utilizado para essa ligação encontra-se no Anexo I (B).

As ligações 02, 03 e 04 levaram em consideração as relações homofílicas identificadas nos dados empíricos e apresentadas na seção 4.1. A ligação 02 foi estabelecida utilizando a variável orientação sexual. Para isso, primeiramente foram calculadas as probabilidades de um homossexual convidar outro homossexual, de um homossexual convidar um bissexual, de um bissexual convidar um homossexual e de

um bissexual convidar outro bissexual, de acordo com os dados empíricos. Em seguida, essas probabilidades foram utilizadas como ponderação na hora de selecionar as ligações entre os elementos. O *script* utilizado para estabelecer essas ligações é apresentado no Anexo I (C).

A ligação 03 por sua vez, foi estabelecida utilizando a variável "idade". Nesse caso, a probabilidade de um indivíduo A conhecer um indivíduo B tem probabilidade calculada a partir dos parâmetros do modelo apresentado na tabela 4.5. Em seguida, essa variável foi utilizada como ponderação para a seleção das ligações. O *script* utilizado para a geração dessa estrutura está detalhado no Anexo I (D).

Finalmente, a ligação 04 foi construída considerando ambas as variáveis, "orientação sexual" e "idade". Primeiramente, como foi verificado que havia independência entre as duas características (resultado não mostrado), obteve-se uma variável combinada, através da multiplicação das duas, que foi utilizada como ponderação na escolha das ligações. Esse procedimento está descrito no Anexo I (E).

A implementação destes quatro algoritmos geradores de redes criou quatro populações distintas, compostas pelos mesmos elementos, porém conectados de forma diferente.

4.2.2. Simulação dos casos infectados.

Uma vez que as quatro estruturas populacionais foram criadas, o passo seguinte foi a determinação dos casos infectados dentre os indivíduos de cada população. Novamente, quatro diferentes modelos de distribuição de casos foram elaborados. Para não confundir com as estruturas de ligação, os modelos de distribuição de infecções foram denominados Infecção A, B, C e D. Em todos os cenários, a prevalência utilizada foi **0,2**, e esse valor foi escolhido arbitrariamente.

O modelo de infecção A consistiu na distribuição aleatória dos casos infectados na população (ver Anexo I (F)).

O modelo B assume que a probabilidade de estar infectado aumenta com o grau da pessoa. Nesse caso, quanto maior o número de contatos de um indivíduo, maior também seria a sua chance de estar infectado. Para esse modelo, os casos foram selecionados utilizando amostragem aleatória ponderada, onde o grau foi a variável de ponderação. O *script* utilizado para a determinação desses casos infectados está apresentado no Anexo I (G).

O modelo infecção C utiliza parâmetros derivados do ajuste do modelo logístico aos dados empíricos (tabela 4.7). Para isso, vale lembrar que o ajuste de um modelo de regressão logística é dado pela seguinte equação:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + X_1\beta_1 + ... + X_n\beta_n = X\beta \qquad (22)$$

Aplicando aos valores apresentados na tabela 4.7, a equação 22 fica da seguinte forma:

$$\log\left(\frac{p}{1-p}\right) = -2{,}28 + X_1 * (0{,}076) + X_2 * (-1{,}195) + X_3 * (-1{,}399) + X_4 * (-0{,}68) + X_5 * (-0{,}111) \qquad (23)$$

A partir disso, e realizando uma manipulação algébrica sobre a equação 22, é possível calcular a probabilidade de um elemento estar infectado, utilizando a seguinte equação:

$$p = \frac{e^{x\beta}}{(1+e^{x\beta})} \qquad (24)$$

Como as informações sobre essas variáveis estavam disponíveis para todos os indivíduos da população, essa probabilidade foi calculada individualmente, e posteriormente, utilizada para selecionar os casos infectados pela infecção C através de amostragem aleatória ponderada, como pode ser observado no Anexo I (H).

Finalmente, o quarto modelo de distribuição de casos, a infecção D, se deu a partir da criação de uma rede de transmissão de infecção. Para isso, 50 indivíduos foram aleatoriamente selecionados e definidos como infectados. Em seguida, os contatos desses indivíduos foram identificados, os quais receberam status de "infectado", com probabilidade **p**. Essa probabilidade de infecção se baseou no produto entre o total de ligações do indivíduo transmissor infectado e um parâmetro arbitrário, escolhido como 0,7. A partir disso, para cada indivíduo infectado, o número de novos casos foi determinado e os novos indivíduos (dentro dos seus contatos), selecionados. Esse processo de transmissão se repetiu até atingir a prevalência pré-estabelecida de 20% . O *script* utilizado para essa situação está apresentado no Anexo I (I).

Dessa forma, foram elaboradas quatro estruturas de ligações entre os elementos da população e quatro estruturas de distribuição dos casos infectados, resultando em dezesseis cenários populacionais diferentes, pois todas as combinações

possíveis entre as estruturas de ligação e infecção foram realizadas. Para referência, a tabela 4.8 resume os cenários investigados.

Tabela 4.8. Cenários investigados.

Cenário	Estrutura de ligação	Distribuição da infecção
1A	Aleatória	Aleatória
2A	Homofílica (OS*)	Aleatória
3A	Homofílica (idade)	Aleatória
4A	Homofílica (OS* + idade)	Aleatória
1B	Aleatória	Dependente do grau
2B	Homofílica (OS*)	Dependente do grau
3B	Homofílica (idade)	Dependente do grau
4B	Homofílica (OS* + idade)	Dependente do grau
1C	Aleatória	Dependente de covariáveis
2C	Homofílica (OS*)	Dependente de covariáveis
3C	Homofílica (idade)	Dependente de covariáveis
4C	Homofílica (OS* + idade)	Dependente de covariáveis
1D	Aleatória	Transmissão
2D	Homofílica (OS*)	Transmissão
3D	Homofílica (idade)	Transmissão
4D	Homofílica (OS* + idade)	Transmissão

*OS = orientação sexual

4.3. Obtenção das amostras geradas por RDS

Depois de obter as dezesseis estruturas populacionais, o passo seguinte consistiu na implementação do processo de amostragem utilizando a técnica RDS. Foi determinado que o tamanho mínimo amostral seria de 500 participantes. Utilizando um nível de confiança de 95%, e a prevalência populacional de 0,2 esse tamanho amostral permite um erro máximo de 0,035.

Outros parâmetros do RDS pré-fixados foram o número de sementes, 5, e o número de convites distribuídos por participante, 3. A escolha das sementes foi realizada de forma aleatória ponderada, considerando para ponderação os graus individuais. Isso porque, segundo é recomendado pela metodologia, é importante que as sementes tenham redes de contato grande, garantindo assim, a continuidade da amostra. Feito isso, dois processos de amostragem foram implementados no R.

Recrutamento completo. No primeiro, cujo *script* pode ser visualizado no Anexo II (A), cada participante selecionado gerou três filhos, ou seja, cada participante gerou três novos participantes para ocupar a onda seguinte. Neste processo, um participante só pôde gerar menos de três filhos se seu grau fosse menor do que três ou se todas as suas ligações já pertencessem à amostra, não tendo esse número de contatos "disponíveis" para ingressar na amostra (a amostragem é sem

reposição). Para diferenciar os dois processos de recrutamento, esse primeiro foi denominado recrutamento completo.

Recrutamento aleatorizado. A segunda forma de recrutamento que foi implementada e cujo *script* está apresentado no Anexo II (B) foi denominada recrutamento aleatorizado. Nesta, o número de indivíduos que esse participante irá recrutar com sucesso é um número aleatório. Como o número máximo de convites distribuídos por participante foi determinado como três, esse valor pôde variar entre zero e três. No entanto, buscando sempre se aproximar o máximo possível de uma situação real, em vez de utilizar probabilidades iguais para essa escolha, essa probabilidade seguiu as mesmas proporções observadas nos dados empíricos. A figura 4.5., apresenta a distribuição de convites bem sucedidos por participante no Projeto "Semear Saúde".

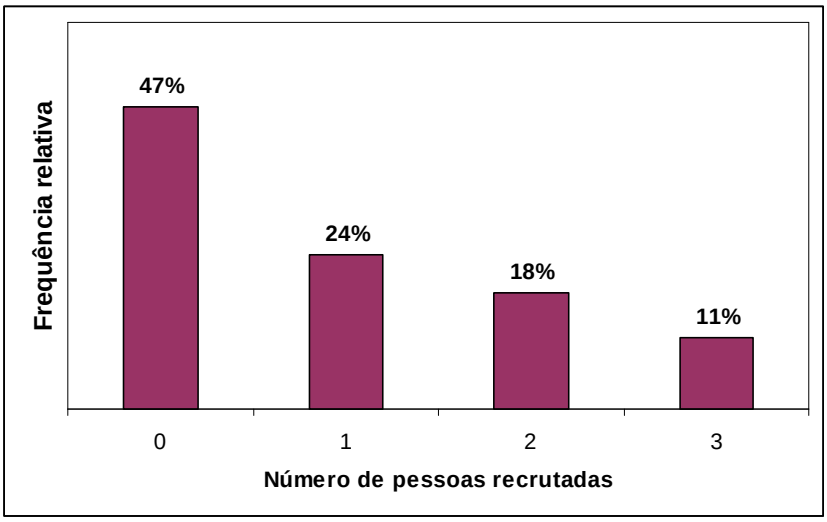

Figura 4.5. Número de pessoas recrutadas com sucesso por participante – dados do Projeto "Semear Saúde".

4.4. Estimação das prevalências amostrais

Com a metodologia apresentada até esse momento, foi possível gerar 16 populações virtuais, e amostras destas populações utilizando o método RDS. O próximo passo foi a obtenção das estimativas de prevalência de infecção nas populações a partir das amostras geradas. Dois métodos para o cálculo dessas estimativas foram implementados em R, e ambos os processos estão detalhados no Anexo III. O primeiro foi denominado **estimativa simples** e consistiu apenas em calcular a proporção de indivíduos infectados e não infectados na amostra. O objetivo

de se calcular essa estimativa simples foi a posterior comparação com as estimativas obtidas pelo modelo de Heckathorn (2002), para avaliar se a correção proposta nesse cálculo traria diferenças significativas para as estimativas.

O segundo método de cálculo foi denominado **RDS** e seguiu o modelo proposto por Heckathorn (2002). Existe um *software* que já realiza esses cálculos, o RDSat (Volz *et al.*, 2007). No entanto, como muitas amostras foram obtidas e para todas elas seria necessário obter as estimativas corrigidas por esse modelo, optou-se por implementar essa rotina no R. Mesmo assim, o RDSat foi utilizado para avaliar se o modelo implementado em R apresentaria os mesmos resultados que o RDSat e isso foi verificado. Diversas amostras foram obtidas e suas estimativas foram calculadas por esses dois programas, apresentando resultados com variações muito pequenas, em geral na terceira casa decimal.

Para avaliar as propriedades estatísticas dos estimadores de prevalência gerados, não seria suficiente observar apenas uma amostra de cada cenário gerado. Dessa forma, o processo de amostragem foi repetido 100 vezes para cada um dos 16 cenários e, para cada um deles, as respectivas estimativas de prevalência foram calculadas pelo método simples e pelo de Heckathorn. Dessas 100 vezes, 50 amostras foram obtidas utilizando o recrutamento completo, e as outras 50, o recrutamento aleatorizado. Análises preliminares sugeriram que 50 amostras seriam suficientes para se obter resultados comparáveis. Em seguida, as estimativas de prevalência foram calculadas, armazenadas e os resultados serão apresentados no próximo capítulo.

5. Resultados

Um dos objetivos desta dissertação é comparar a performance de dois estimadores de prevalência, o simples e o RDS, em amostras geradas por cadeias de referência em populações com diferentes padrões de agregação social e distribuição de casos positivos. Os resultados serão apresentados na forma de medidas resumo e com a elaboração de gráficos. Na seção 5.1., são apresentados os resultados para os cenários onde todos os participantes buscam gerar três filhos, denominado *recrutamento completo*. Relembrando, nesses casos, um participante gerou menos de três filhos apenas quando seu número de conhecidos que ainda não pertenciam à amostra era menor do que três. Na seção 5.2., são apresentados os resultados para as simulações que randomizaram o número de filhos recrutados (de zero a três), denominado *recrutamento aleatorizado*.

Os gráficos foram divididos de acordo com as quatro estruturas de distribuição de casos explicadas no capítulo 4. Vale lembrar que em todas as simulações, a prevalência utilizada foi de **0,2**.

Visando estabelecer comparações entre os resultados apresentados nas seções 5.1 e 5.2 deve-se atentar para o fato de que os eixos dos gráficos são diferentes. Na primeira seção, a amplitude utilizada foi [0 ; 0,4], enquanto, na segunda, a amplitude foi [0; 1]. A razão de não escolher intervalos iguais foi que, ao utilizar o segundo intervalo para os primeiros cenários, as figuras ficaram muito condensadas, não sendo possível observar as diferenças entre eles. Isso mostrou que a variabilidade nas estimativas é menor quando todos os participantes conseguem recrutar o número devido de filhos. Outro fato a ser considerado é que o tamanho da amostra varia muito quando o número de filhos é escolhido aleatoriamente, como também será discutido a seguir.

5.1. Recrutamento completo.

Como já foi citado, o processo de amostragem deveria seguir até atingir um tamanho mínimo de 500 elementos. Assim, ao realizar o recrutamento de exatamente três filhos, todas as amostras ficaram com tamanhos parecidos, em torno de 580 participantes. Pequenas variações se deve ao fato de que alguns indivíduos não possuem três filhos disponíveis para recrutamento (seja porque o grau era menor do que três ou porque as ligações já pertenciam à amostra). Cinco sementes foram selecionadas para o início do recrutamento, o que fez com que o tamanho de amostra escolhido fosse

atingido em apenas quatro ondas. Esse número de ondas é pequeno, em comparação com as recomendações teóricas (Gile & Handcock, 2009). A figura 5.1. traz um exemplo de cadeias geradas utilizando recrutamento completo. Os gráficos de rede apresentados nessa dissertação foram feitos com a utilização do *software* NetDraw, versão 2.084 (Borgatti, 2009). Cada círculo na figura representa um indivíduo pertencente à amostra e os círculos maiores representam as sementes. Como é possível observar, as cinco cadeias são muito parecidas, o que acontece devido ao recrutamento completo.

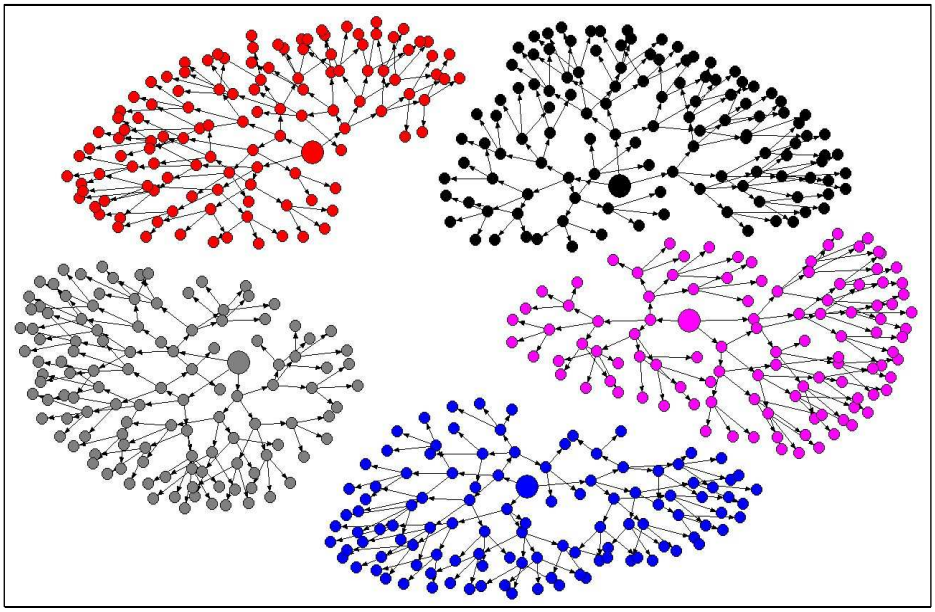

Figura 5.1. Exemplo de amostra gerada utilizando recrutamento completo.

A tabela 5.1. apresenta a prevalência mediana estimada pelos métodos simples e RDS, para os dados gerados por recrutamento completo, considerando os 16 cenários de infecção e ligação. Com não foi testada nenhuma hipótese sobre a distribuição de probabilidades das estimativas, optou-se por utilizar as medidas de mediana e amplitude (cálculos da prevalência média também foram realizados, mas não serão apresentados, pois pouca diferença foi observada entre a média e a mediana, mostrando que a distribuição das estimativas está mais ou menos simétrica em torno das medidas centrais).

Pela tabela 5.1., pode-se observar que o modelo de Heckathorn (2002) apresenta bons resultados, pois a mediana para as estimativas está sempre bem próxima à prevalência verdadeira de 0,2. Além disso, observa-se a importância da ponderação deste, principalmente, nos cenários nos quais o status de infecção está associado ao

grau dos indivíduos (infecção B) pois ao comparar as estimativas RDS e simples desses cenários, é possível perceber que há uma tendência de superestimação da prevalência quando o método RDS não é utilizado. Essa vantagem do modelo de Heckathorn também é verificada nos cenários onde os indivíduos infectados foram gerados por contágio, formando clusters (infecção D), pois é possível verificar que as estimativas RDS se concentram mais em torno da prevalência real do que as estimativas simples, embora ambos os casos se aproximem de 0,2.

Tabela 5.1. Medidas resumo da estimativa de prevalência calculada por amostragem RDS, utilizando o recrutamento completo.

		Mediana		Amplitude	
		Simples	RDS	Simples	RDS
Infecção A	Ligação 01	0,204	0,205	0,081	0,108
	Ligação 02	0,197	0,190	0,063	0,116
	Ligação 03	0,193	0,192	0,066	0,100
	Ligação 04	0,198	0,198	0,082	0,133
Infecção B	Ligação 01	0,329	0,192	0,090	0,098
	Ligação 02	0,333	0,201	0,077	0,069
	Ligação 03	0,329	0,194	0,081	0,085
	Ligação 04	0,338	0,202	0,088	0,068
Infecção C	Ligação 01	0,196	0,198	0,093	0,121
	Ligação 02	0,199	0,208	0,053	0,100
	Ligação 03	0,200	0,193	0,080	0,136
	Ligação 04	0,201	0,199	0,067	0,128
Infecção D	Ligação 01	0,222	0,192	0,067	0,093
	Ligação 02	0,210	0,202	0,091	0,096
	Ligação 03	0,219	0,200	0,081	0,098
	Ligação 04	0,230	0,209	0,068	0,114

Para verificar se as diferenças observadas entre os dois métodos de estimação são estatisticamente significativas, um teste de Wilcoxon para diferença de medianas foi aplicado. Os resultados podem ser vistos na tabela 5.2.. Por essa tabela, observa-se que as medianas dos dois métodos foram significativamente diferentes nos cenários das infecções B e D. Além disso, é observada também uma diferença significativa para o cenário 02 (Infecção A/ Ligação 02).

Tabela 5.2. Teste de Wilcoxon para diferença de medianas entre as estimativas Simples e RDS no recrutamento completo.

		Estatística (W)	P-valor
Infecção A	Ligação 01	1354	0,4755
	Ligação 02	981	0,06416
	Ligação 03	1131	0,4139
	Ligação 04	1178	0,6221
Infecção B	Ligação 01	0	< 0,01
	Ligação 02	0	< 0,01
	Ligação 03	0	< 0,01
	Ligação 04	0	< 0,01
Infecção C	Ligação 01	1301	0,7277
	Ligação 02	1353	0,4797
	Ligação 03	1134	0,4259
	Ligação 04	1182	0,6417
Infecção D	Ligação 01	430	< 0,01
	Ligação 02	870	< 0,01
	Ligação 03	861	< 0,01
	Ligação 04	788	< 0,01

Ao analisar a variabilidade das estimativas nas 50 repetições realizadas para cada cenário, foi observado que, de forma geral, as estimativas RDS apresentam maior variabilidade do que as estimativas simples, embora essa diferença não seja muito grande. As figuras 5.2., 5.3., 5.4. e 5.5. trazem os resultados gráficos para as estimativas de prevalência, que foram obtidos com a construção de *boxplots*.

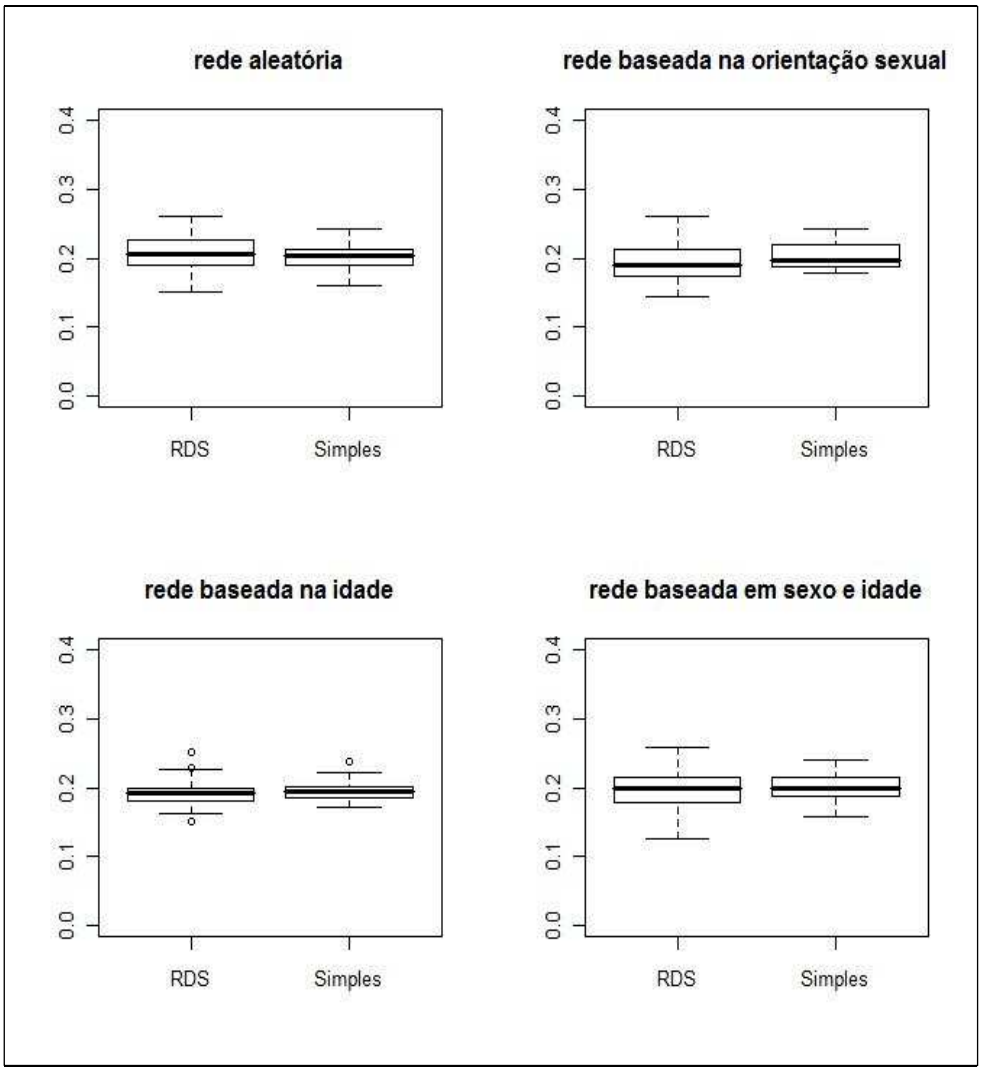

Figura 5.2. Box-plots das estimativas de prevalência obtidas por recrutamento completo, quando a distribuição de pessoas infectadas na população é aleatória simples (cenários 1A, 2A, 3A e 4A da tabela 4.8.).

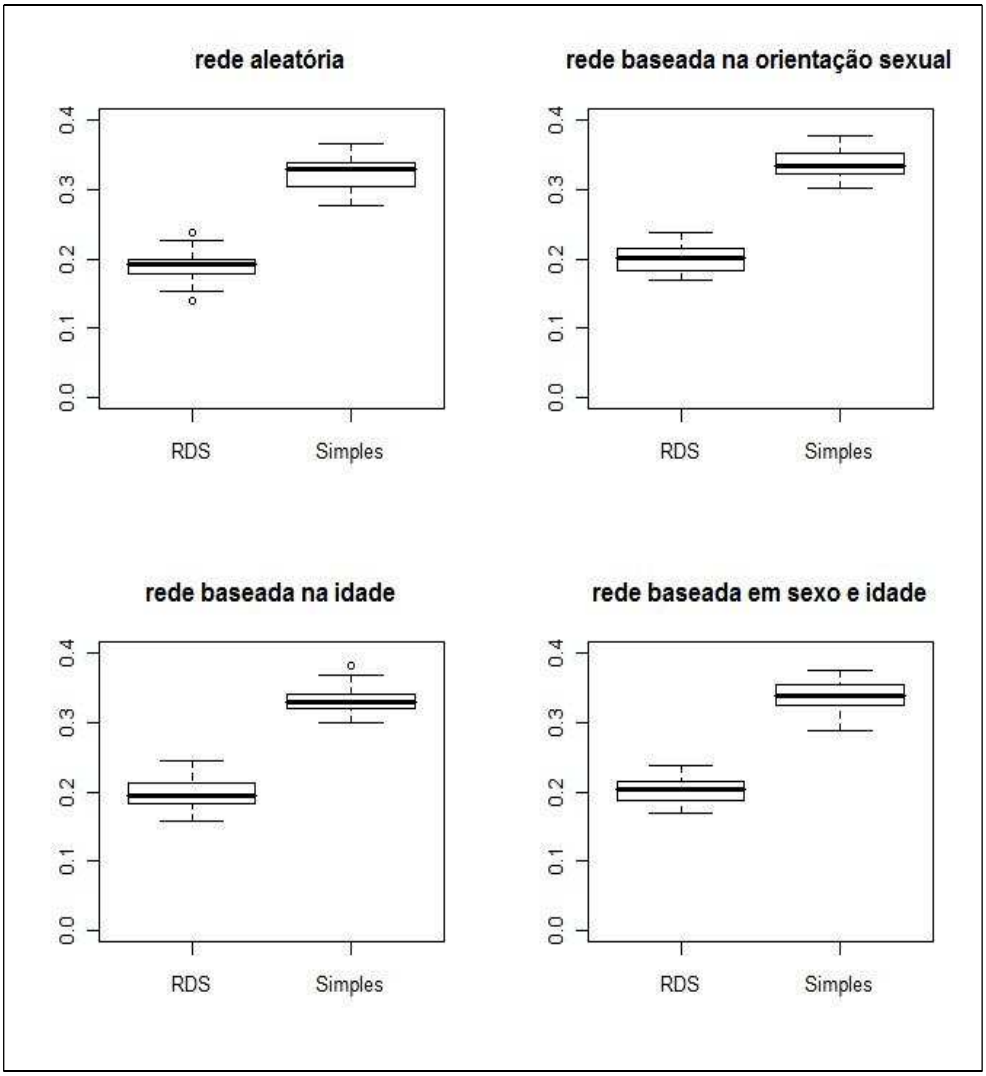

Figura 5.3. Boxplots das estimativas de prevalência obtidas por recrutamento completo, quando a distribuição de pessoas infectadas na população é aleatória ponderada, com probabilidade de seleção proporcional ao grau (cenários 1B, 2B, 3B e 4B da tabela 4.8.).

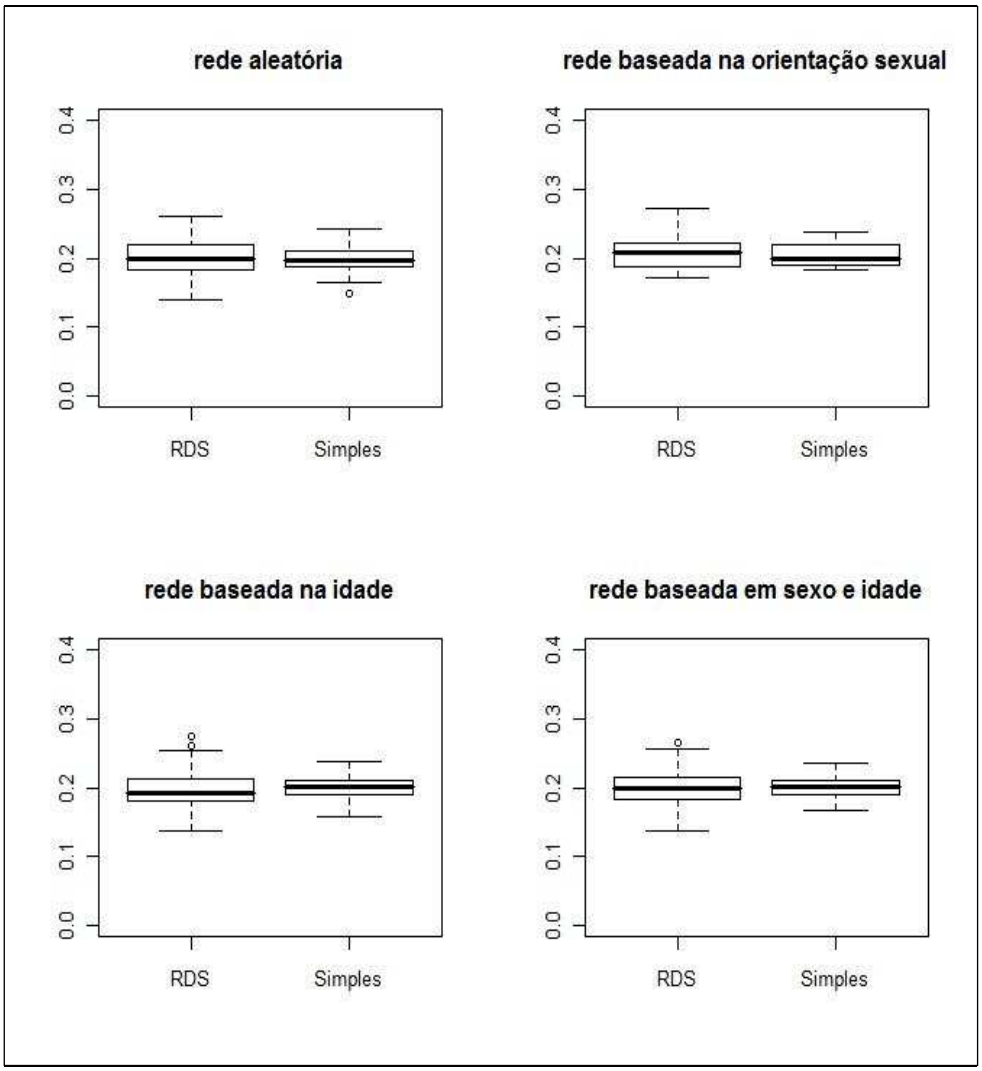

Figura 5.4. Boxplots das estimativas de prevalência obtidas por recrutamento completo, quando a distribuição de pessoas infectadas na população é aleatória ponderada, com probabilidade de infecção determinada por covariáveis de determinação do risco associado (cenários 1C, 2C, 3C e 4C da tabela 4.8.).

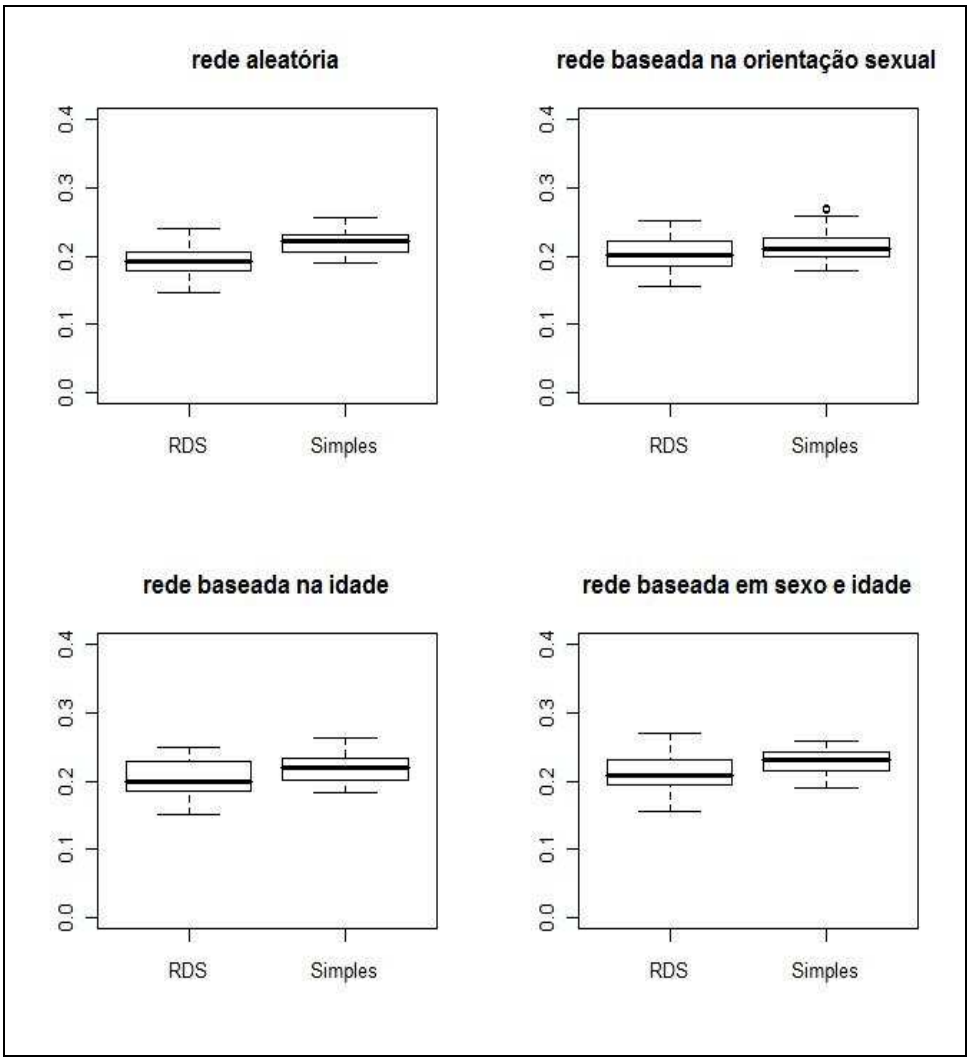

Figura 5.5. Boxplots das estimativas de prevalência obtidas por recrutamento completo, quando a distribuição de pessoas infectadas na população é realizada por cadeia de transmissão (cenários 1D, 2D, 3D e 4D da tabela 4.8.).

Além disso, foi realizada uma observação visual dos indivíduos amostrados em relação ao status para a infecção. O objetivo era identificar, pelo menos de forma visual, se a forma como as pessoas estão conectadas na população tinha alguma relação com a distribuição espacial dos casos infectados. No entanto, não foi possível observar diferenças entre as distribuições. A figura 5.6. apresenta exemplos de amostras geradas para a infecção D, onde cada gráfico se refere a um tipo de ligação e os indivíduos infectados estão apresentados em quadrados pretos, e onde nenhuma estrutura específica se destaca. Vale lembrar ainda que, como não foi observada diferença significativa entre as estruturas de ligação, escolheu-se a construção desses gráficos apenas para os cenários onde a infecção foi introduzida em clusters. Essa escolha se baseou no fato

desse ser o cenário mais próximo às principais investigações de interesse que utilizam o método RDS, ou seja, as doenças sexualmente transmissíveis.

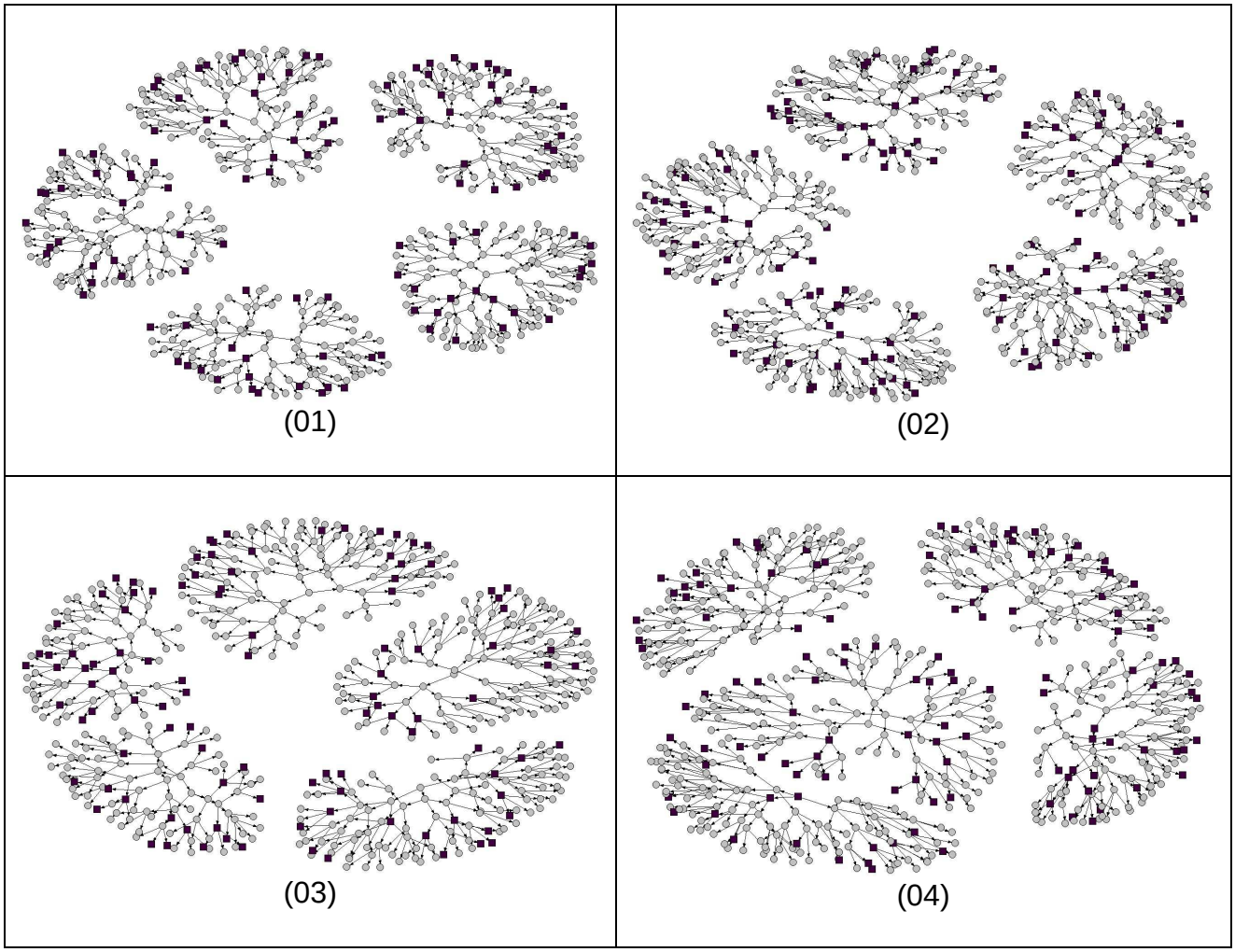

Figura 5.6. Representação gráfica de indivíduos infectados, partindo de amostras de cadeias de recrutamento completo e diferentes tipos de ligação entre os indivídudos. (01) Ligações aleatórias; (02) Ponderadas pela orientação sexual; (03) Ponderada pela idade; (04) Ponderada por orientação sexual e idade.

5.2 Recrutamento aleatorizado.

A primeira consideração feita no cenário de recrutamento aleatorizado se refere ao tamanho das amostras. Foi observada uma variação muito grande nos tamanhos das amostras geradas ao repetir-se o processo de amostragem RDS, onde a menor amostra teve apenas 5 participantes – pois as sementes não deram frutos – e a maior, 519. Para ilustrar essas situações, a figura 5.7 apresenta o gráfico de duas das cadeias de recrutamento geradas, uma bastante longa, que atinge o número pré-determinado de participantes, e outra menor, com poucos participantes.

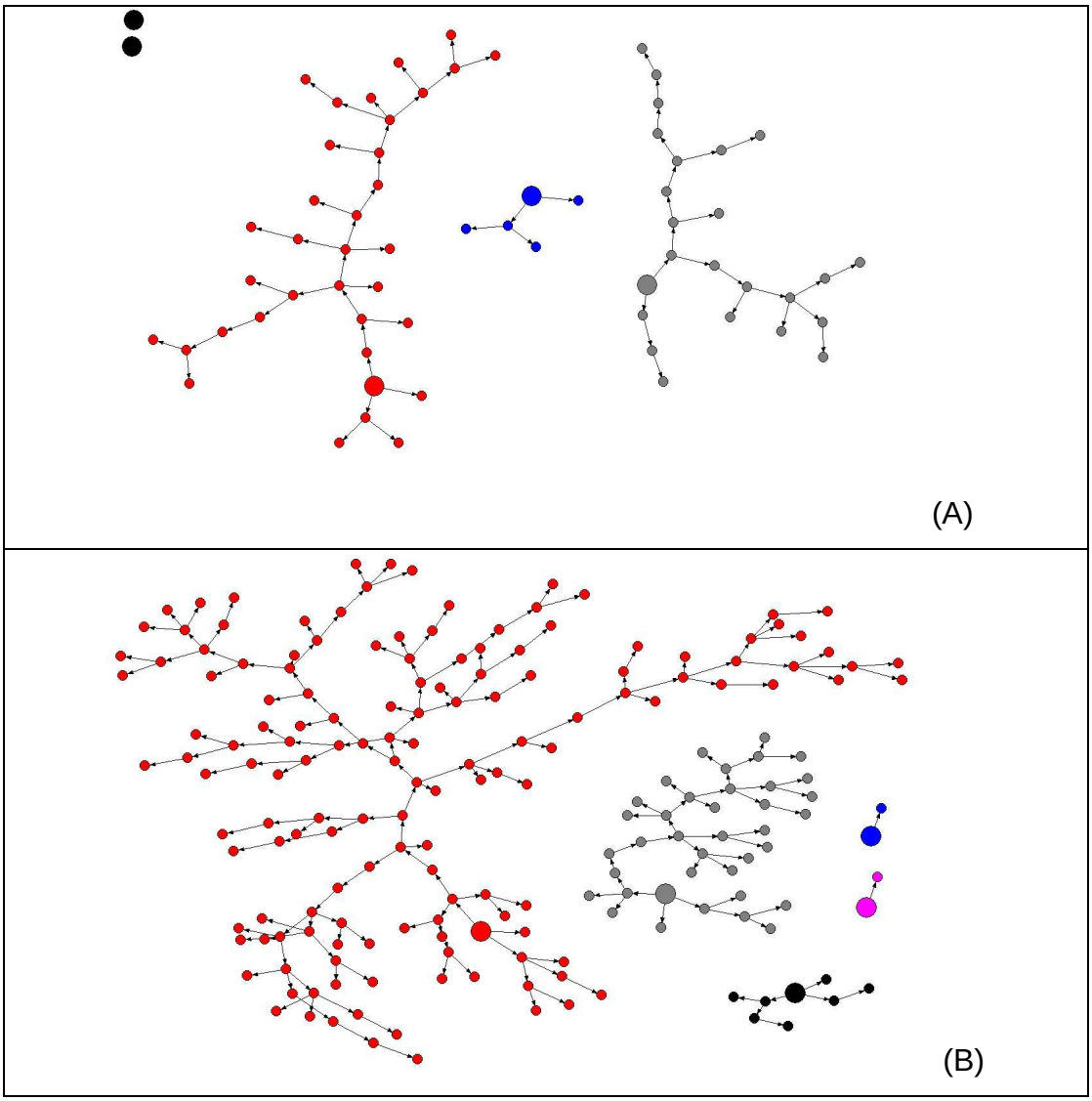

Figura 5.7. Exemplos de amostras utilizando recrutamento aleatorizado, com (A) poucos participantes e (B) muitos participantes.

A figura 5.7. traz também outra informação relevante, que se refere ao número máximo de ondas que cada amostra conseguiu atingir. Como nesse processo o número

de filhos gerados é escolhido aleatoriamente, o número de ondas também teve grande variação. Para amostras onde as sementes não frutificaram, o número de ondas foi zero, enquanto que para as amostras que se aproximaram do tamanho mínimo desejado, até 93 ondas foram observadas. Ao se pensar num processo empírico de amostragem, uma alternativa para os casos onde a amostra não atinge o tamanho necessário é a inclusão de novas sementes, ou o estímulo dos participantes já envolvidos.

Além do número de ondas, é importante também fazer algumas considerações sobre o tamanho da amostra. Intuitivamente, é possível se pensar que, quanto maior o tamanho da amostra, melhor também será a estimativa da prevalência da característica de interesse. As figuras 5.8., 5.9., 5.10 e 5.11. apresentam as estimativas obtidas por ambos os métodos, em relação ao tamanho da amostra. Nessas figuras, é possível comprovar que realmente, embora algumas estimativas calculadas a partir de amostras pequenas estejam com valores próximos a 0,2, a medida que o tamanho da amostra aumenta, as estimativas vão se aproximando cada vez mais desse valor. Adicionalmente, vale destacar que esse comportamento foi observado para todas as situações geradas, não sendo influenciado, pelo menos aparentemente, pelos diferentes cenários.

Uma diferença observada nesses gráficos se refere aos cenários onde a estrutura de ligação entre as pessoas foi simulada por cadeias. Como é possível verificar, enquanto os outros cenários tiveram amostras com tamanhos próximos a 500, nesses cenários, as amostras dificilmente passaram de 200. Uma possível explicação para isso está na forma de ligação entre as pessoas. Porém, não foram realizadas investigações mais profundas sobre essa questão. Em contrapartida, nesses cenários, mesmo amostras com tamanhos pequenos geraram estimativas próximas a 0,2. Isso reflete mais uma vez a importância de se considerar o método proposto de ponderação e indica também o bom desempenho do processo de amostragem.

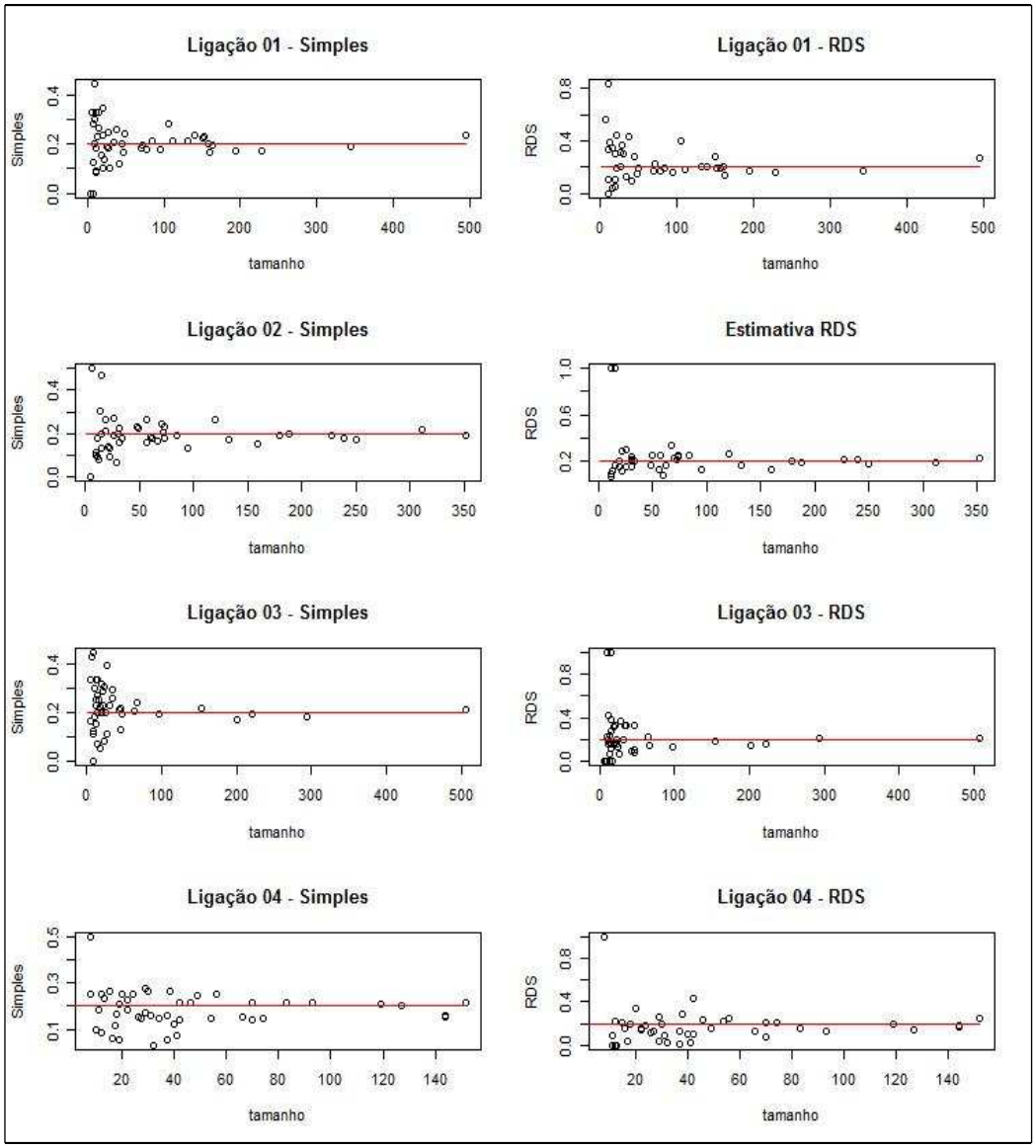

Figura 5.8. Efeito do tamanho final da amostra nas estimativas de prevalência, no cenário de distribuição aleatória de infectados (infecção A).

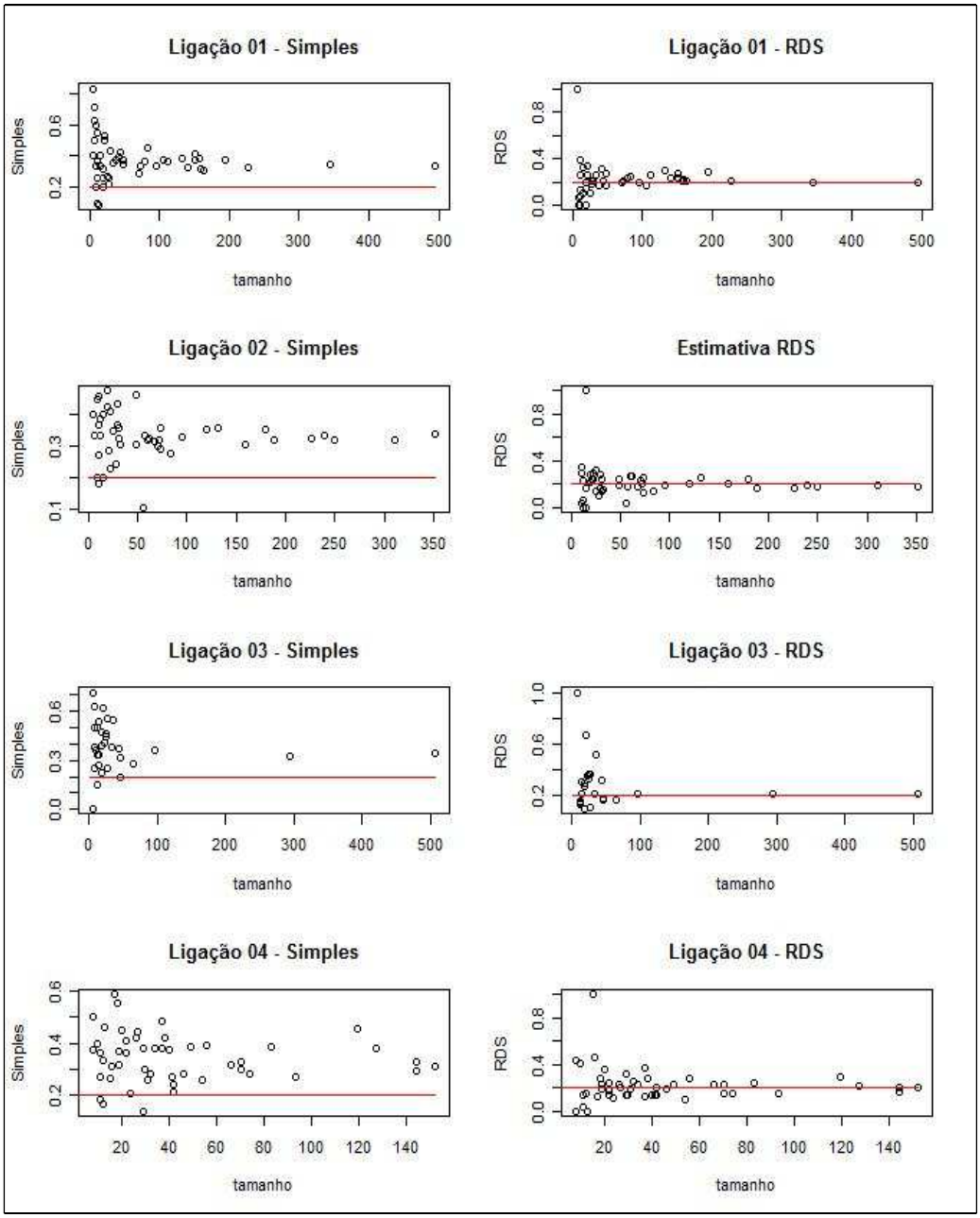

Figura 5.9. Efeito do tamanho final da amostra nas estimativas de prevalência, no cenário de distribuição aleatória de infectados (infecção B).

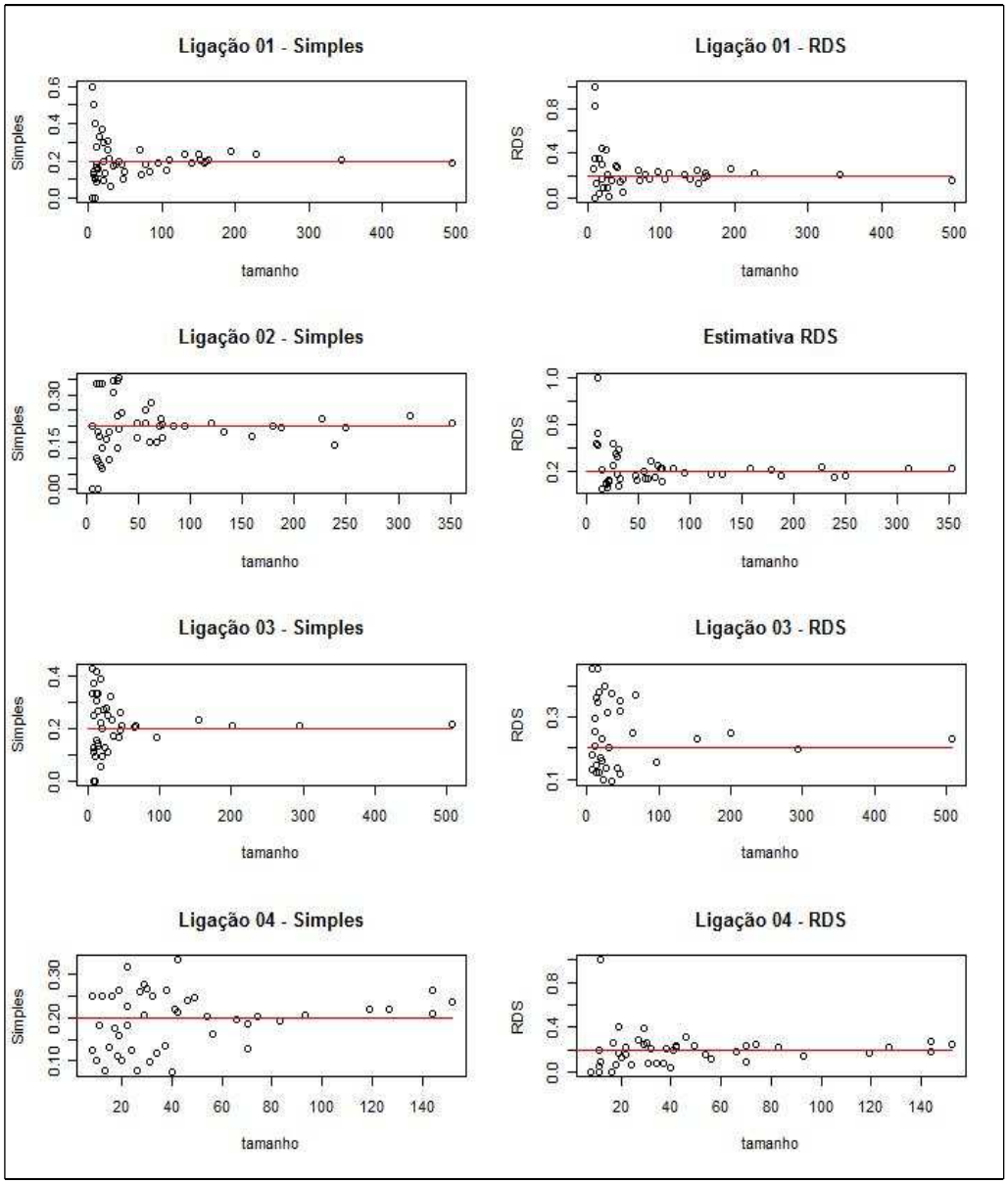

Figura 5.10. Efeito do tamanho final da amostra nas estimativas de prevalência, no cenário de distribuição aleatória de infectados (infecção C).

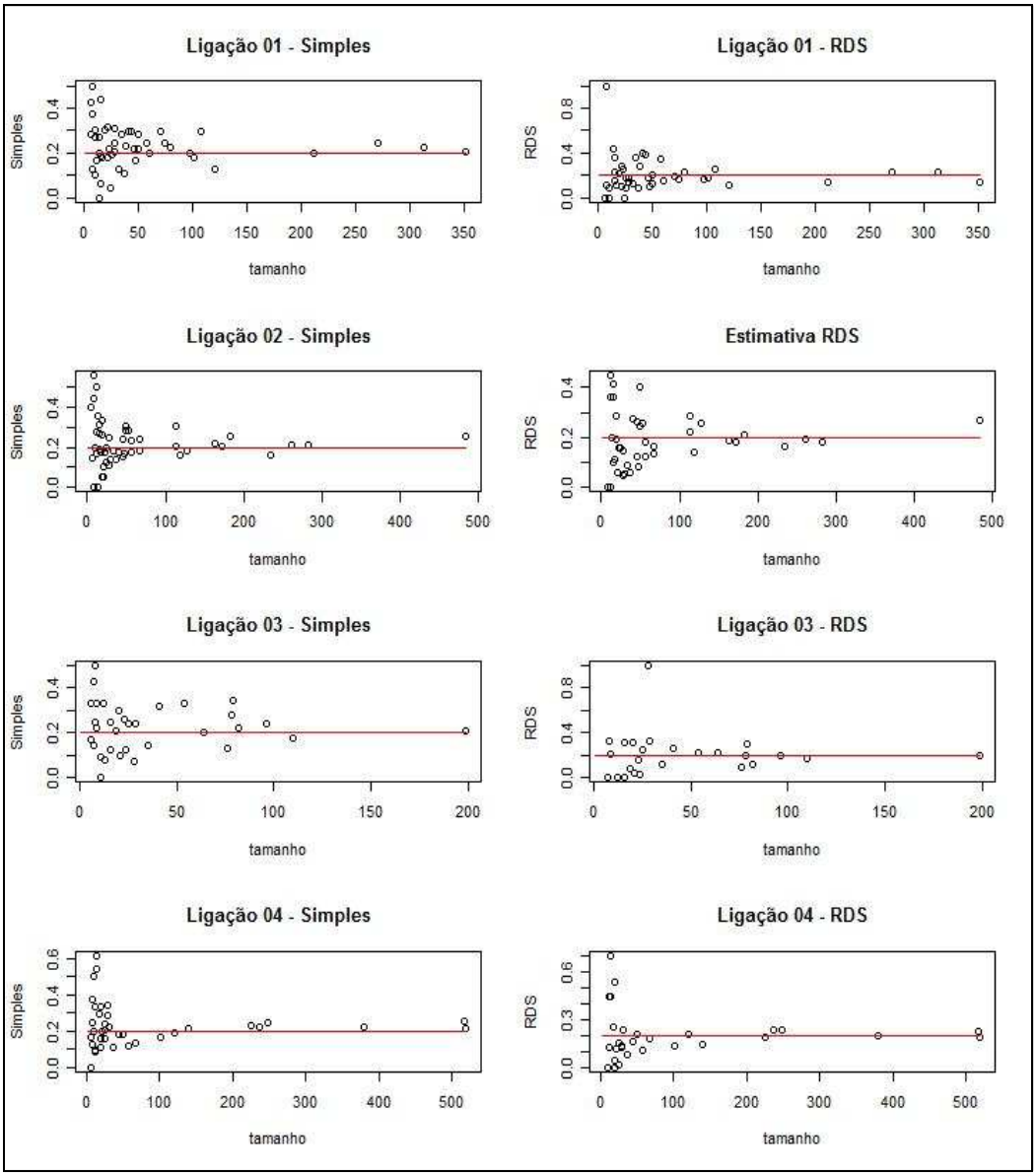

Figura 5.11. Efeito do tamanho final da amostra nas estimativas de prevalência, no cenário de distribuição aleatória de infectados (infecção D).

Para os cenários gerados a partir do recrutamento aleatorizado, também foram extraídas algumas medidas resumo que estão apresentadas na tabela 5.3.. Por essa tabela, é possível perceber que a variabilidade gerada com esse recrutamento é alta e muito maior do que a variabilidade encontrada no recrutamento completo. A mediana das estimativas também se distancia da prevalência verdadeira gerada (0,2) em mais cenários, e com diferenças maiores quando comparadas aos primeiros resultados. Além disso, nota-se novamente a importância de considerar o método proposto para estimar as prevalências, principalmente nos cenários onde os indivíduos infectados foram selecionados de acordo com seus graus (Infecção B).

Tabela 5.3. Medidas resumo para o recrutamento aleatorizado.

		Mediana		Amplitude	
		Simples	RDS	Simples	RDS
Infecção A	Ligação 01	0.200	0.197	0.444	0.829
	Ligação 02	0.183	0.195	0.500	0.943
	Ligação 03	0.210	0.173	0.444	1.000
	Ligação 04	0.182	0.153	0.469	1.000
Infecção B	Ligação 01	0.356	0.208	0.756	1.000
	Ligação 02	0.333	0.195	0.367	1.000
	Ligação 03	0.379	0.265	0.714	0.915
	Ligação 04	0.348	0.201	0.450	1.000
Infecção C	Ligação 01	0.188	0.210	0.600	1.000
	Ligação 02	0.196	0.190	0.355	0.950
	Ligação 03	0.212	0.229	0.429	0.363
	Ligação 04	0.203	0.187	0.258	1.000
Infecção D	Ligação 01	0.219	0.172	0.500	1.000
	Ligação 02	0.202	0.181	0.556	0.448
	Ligação 03	0.221	0.200	0.500	1.000
	Ligação 04	0.211	0.171	0.615	0.700

Para verificar se as diferenças obtidas são estatisticamente significativas, foram realizados testes de Wilcoxon para a diferença entre medianas do método simples e de Heckathorn. Os resultados são apresentados na tabela 5.4., e mostram que diferente da situação do recrutamento completo, nesse caso, algumas estruturas da infecção D não apresentaram diferenças estatisticamente significativas (ligações 02 e 03).

Tabela 5.4. Teste de Wilcoxon para diferença de medianas entre as estimativas Simples e RDS no recrutamento aleatorizado.

		Estatística (W)	P-valor
Infecção A	Ligação 01	1085	0,635
	Ligação 02	1118	0,4606
	Ligação 03	877	0,1762
	Ligação 04	847	0,03858
Infecção B	Ligação 01	319	< 00,01
	Ligação 02	231	< 00,01
	Ligação 03	499	< 00,01
	Ligação 04	392	< 00,01
Infecção C	Ligação 01	1090	0,4674
	Ligação 02	1131	0,6689
	Ligação 03	1186	0,2881
	Ligação 04	968	0,319
Infecção D	Ligação 01	803	0,02464
	Ligação 02	874	0,1689
	Ligação 03	807.5	0,3150
	Ligação 04	747.5	0,0407

As figuras 5.12., 5.13., 5.14 e 5.15. trazem os resultados gráficos para as estimativas de prevalência, que foram obtidos com a construção de *boxplots*. Essas figuras confirmam o que está apresentado anteriormente.

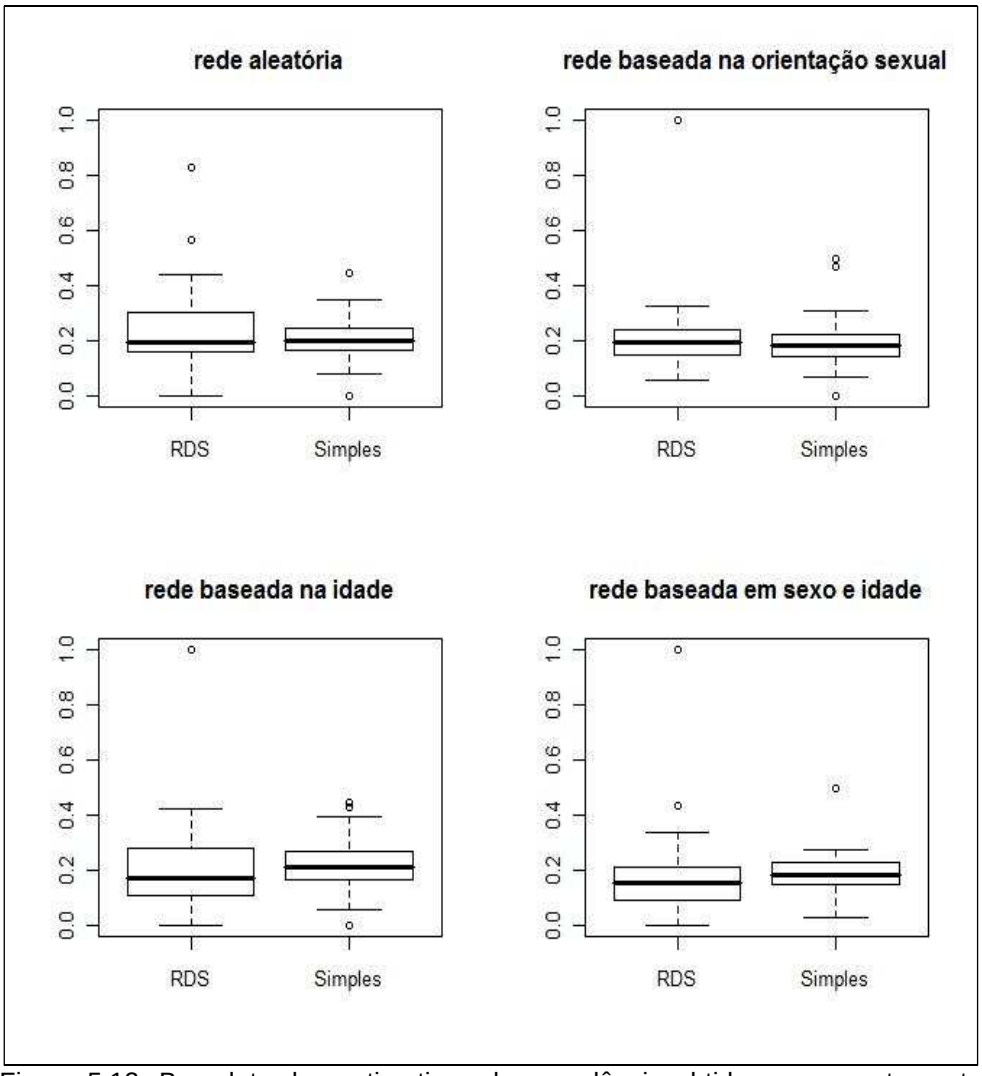

Figura 5.12. Box-plots das estimativas de prevalência obtidas por recrutamento aleatorizado, quando a distribuição de pessoas infectadas na população é aleatória simples (cenários 1A, 2A, 3A e 4A da tabela 4.8.).

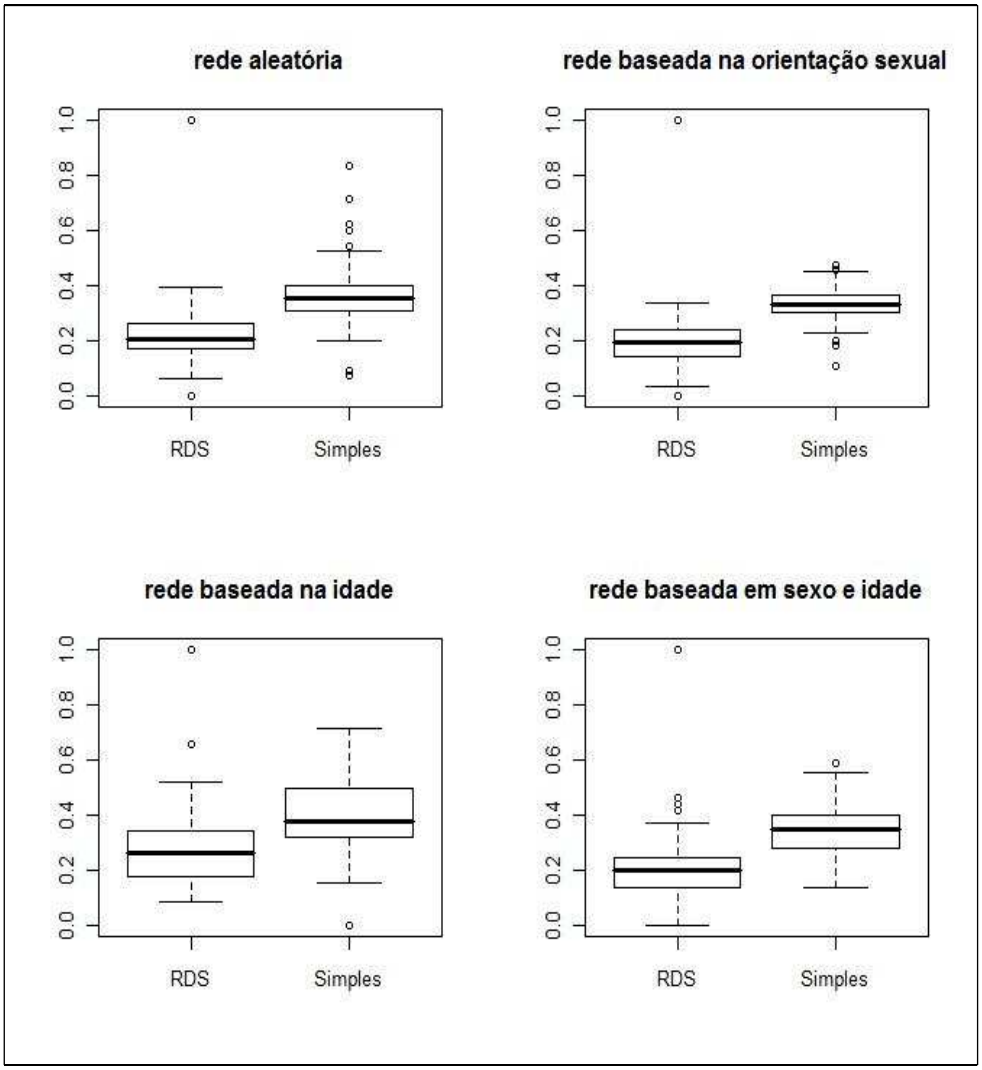

Figura 5.13. Boxplots das estimativas de prevalência obtidas por recrutamento aleatorizado, quando a distribuição de pessoas infectadas na população é aleatória ponderada, com probabilidade de seleção proporcional ao grau (cenários 1B, 2B, 3B e 4B da tabela 4.8.).

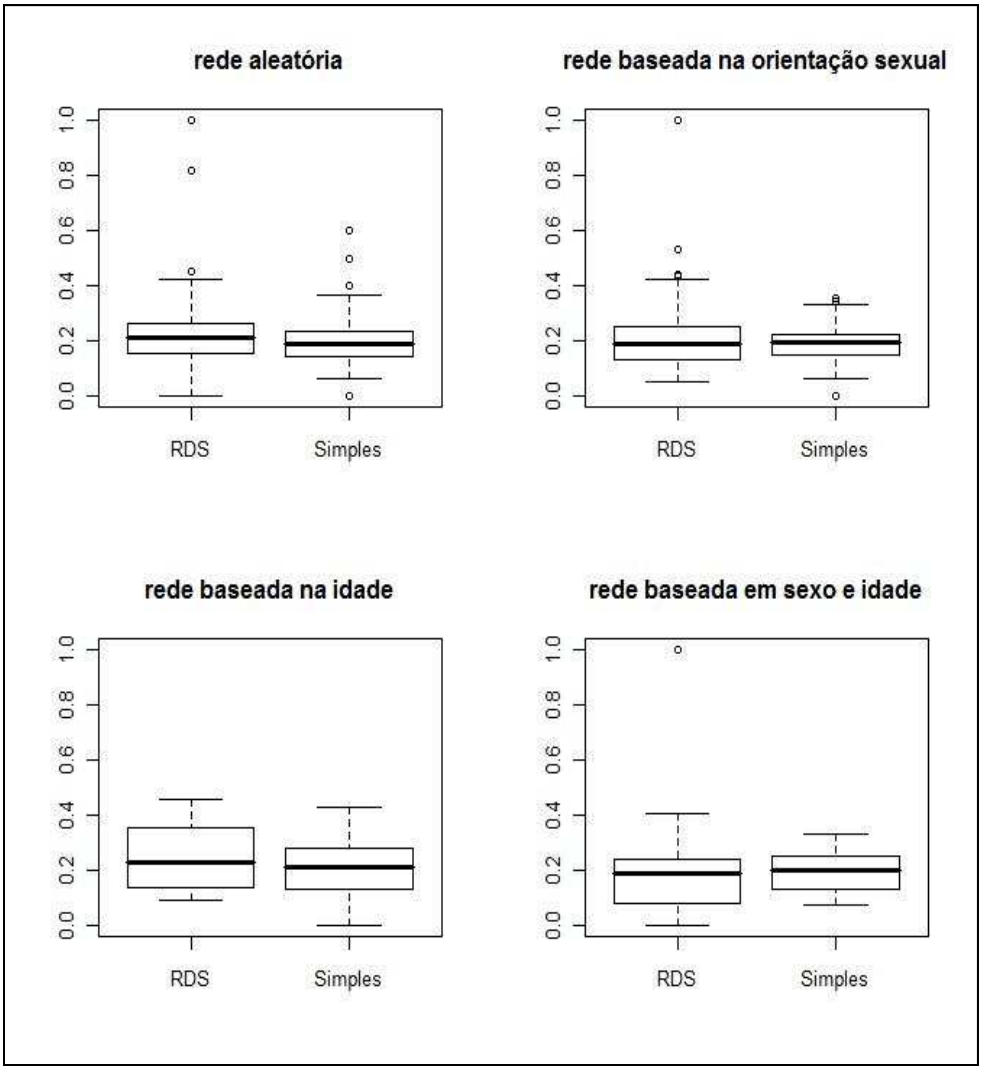

Figura 5.14. Boxplots das estimativas de prevalência obtidas por recrutamento aleatorizado, quando a distribuição de pessoas infectadas na população é aleatória ponderada, com probabilidade de infecção determinada por covariáveis de determinação do risco associado (cenários 1C, 2C, 3C e 4C da tabela 4.8.).

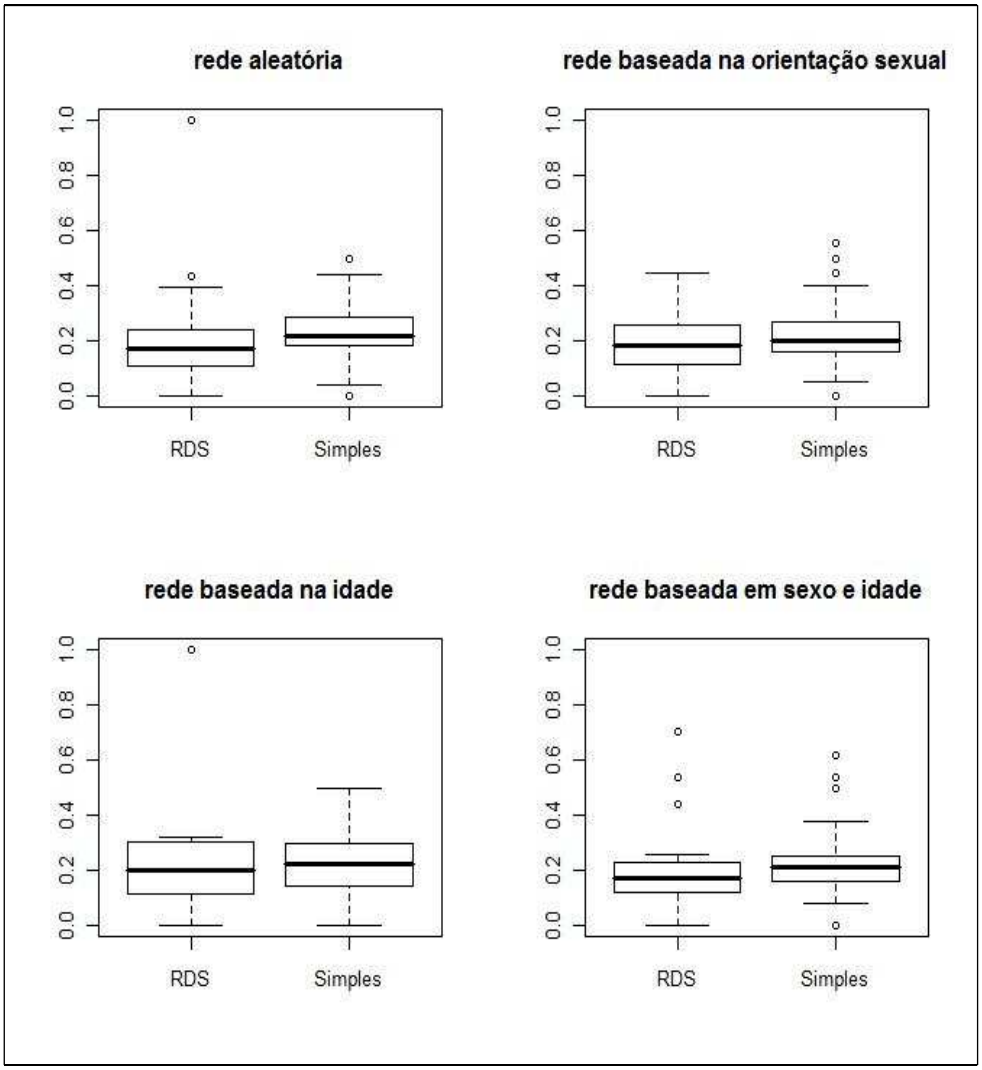

Figura 5.15. Boxplots das estimativas de prevalência obtidas por recrutamento aleatorizado, quando a distribuição de pessoas infectadas na população é realizada por cadeia de transmissão (cenários 1D, 2D, 3D e 4D da tabela 4.8.).

Uma última observação seria sobre a estrutura espacial dos casos de infecção na amostra. No entanto, mais uma vez, não foi possível identificar nenhuma relação entre os indivíduos amostrados e por isso, a figura 5.16. apresenta apenas dois exemplos de cadeias. Vale lembrar que essa ausência de relação era esperada, pois essa variável de infecção não foi considerada, nem no momento de criação das estruturas populacionais, nem no momento de obtenção da amostra, como, por exemplo, Gile e Handcock (2009) fizeram.

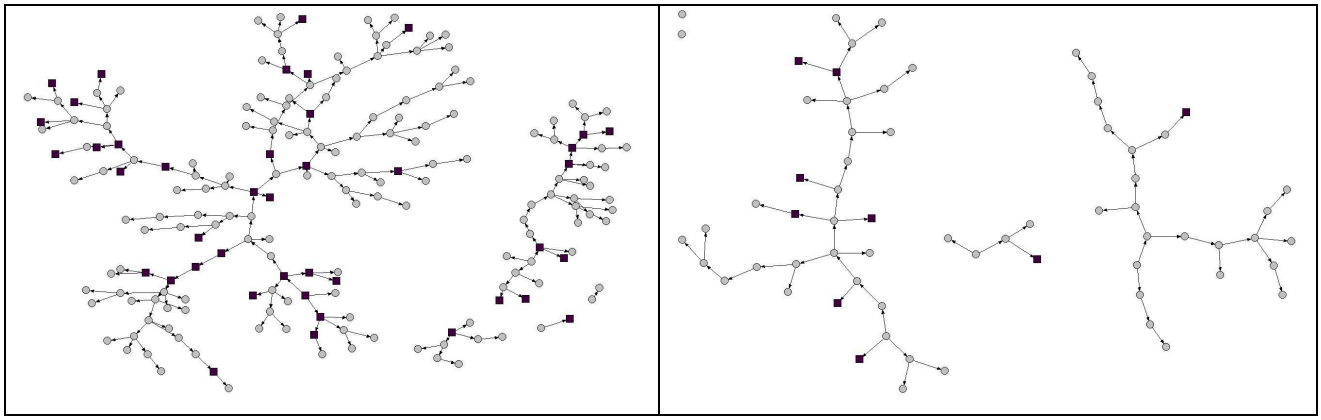

Figura 5.16. Representação gráfica de indivíduos infectados, partindo de amostras de cadeias de recrutamento aleatorizado.

6. Discussão, conclusão e trabalhos futuros.

Essa dissertação teve por objetivo avaliar estimativas de prevalência geradas a partir de amostras obtidas com a utilização da metodologia RDS, considerando diferentes estruturas populacionais. Para isso, foram geradas dezesseis populações distintas, com diferentes formas de conexão entre seus indivíduos e de espalhamento de uma infecção. Para cada uma, foram obtidas cem amostras utilizando RDS. Metade dessas amostras contou com o recrutamento bem sucedido de três outros participantes (recrutamento completo), ao passo que para as outras cinqüenta amostras, antes de selecionar os próximos participantes, foram selecionadas também quantas pessoas seriam recrutadas com sucesso (recrutamento aleatório). Este último é mais compatível com o que se observa em situações reais.

O algoritmo proposto, todo implementado em R, é facilmente adaptável para teste de outras implementações de amostragens baseadas em bola-de-neve, assim como simulação de populações com características diferentes das propostas nesta dissertação.

Com os resultados encontrados, é possível realizar uma avaliação tanto do RDS como forma de recrutamento, como o modelo proposto por Heckathorn para a ponderação e estimação de prevalências. Basicamente, três aspectos podem ser considerados nessa avaliação: 1. o tempo necessário para concluir a amostragem; 2. a precisão das estimativas obtidas, independente da ponderação, ou seja, avaliando sob o olhar de que a metodologia funciona ou não; e 3. o método de ponderação.

Em relação ao tempo necessário para concluir a amostragem, tem-se que, na ocorrência de um recrutamento bem sucedido, como é o esperado na proposta da metodologia, e representado no caso do recrutamento completo, esse processo se dá de forma rápida, incluindo um número pequeno de ondas. Nesse sentido, considerando a escolha aleatória das sementes, a metodologia cumpre o pressuposto de que o crescimento da amostra se dá em taxas geométricas, como foi apresentado na seção 2.4.. Além disso, ainda que com poucas ondas, observou-se que a distribuição das estimativas de prevalência mostrou estimar de forma correta a prevalência populacional, ficando em torno de 20%.

Com relação ao efeito do número de ondas, Gile & Handcock (2009) compararam o comportamento de estimativas geradas também por simulação considerando amostras com 4 ondas e amostras com 6 ondas. Para os casos onde as sementes foram escolhidas de forma aleatória, como ocorreu também nessa dissertação,

não houve diferença significativa entre as distribuições das estimativas de prevalência, sendo que os resultados apresentaram distribuição em torno da prevalência populacional determinada, que também foi de 0,2.

Por outro lado, ao considerar o recrutamento aleatório, os resultados encontrados sugerem que o tempo até atingir o tamanho amostral desejado pode se dar de forma lenta, e inclusive pode não atingir o número mínimo de elementos, determinado para a amostra. Nesse sentido, é importante destacar que no recrutamento aleatório aqui implementado, foram utilizadas probabilidades de seleção do número de participantes que tiveram por base o estudo empírico. Além disso, como é sabido, cada população possui comportamentos diferentes e reage de formas diferentes em relação à aceitação em participar de pesquisas, por exemplo, e dessa forma, para se obter conclusões mais detalhadas sobre a velocidade do processo de recrutamento, seria necessário obter amostras utilizando diferentes probabilidades de seleção. Essa é uma das propostas para trabalhos futuros que podem ser desenvolvidos, onde a idéia é a utilização de outros dados empíricos, de outros projetos que utilizaram ou estão utilizando essa metodologia. Dessa forma, será possível verificar, por exemplo, como se dá o comportamento de populações de usuários de drogas injetáveis, trabalhadores do sexo, ou até mesmo outra comunidade de HSH, sobre essa questão.

Com relação à precisão das estimativas obtidas, é importante observar as diferenças entre o recrutamento completo e o recrutamento aleatório, embora as medianas em ambos os casos se aproximem do valor verdadeiro (0,2) – exceto para a infecção B, quando o espalhamento da doença se deu com probabilidades maiores para pessoas com mais contatos. Essa importância acontece porque na prática, cada estudo parte de apenas uma amostra, e não várias, como foi simulado. Quando o recrutamento completo acontece, há uma variação pequena entre as estimativas observadas, o que é bom e sinaliza que a metodologia parece atingir o objetivo de gerar boas estimativas, ou seja, estimativas fidedignas. Por outro lado, situações de recrutamento aleatório são verificadas com maior freqüência nos estudos empíricos. E para essas situações, é importante destacar a importância de obter amostras com tamanhos razoáveis, pois foi verificado que, quanto maior o tamanho da amostra, mais próximo ao valor verdadeiro populacional as estimativas se aproximaram.

Nesse sentido, dois estudos podem ser citados, para permitir algumas comparações. Um dele é o de Salganik (2006), cujo algoritmo foi apresentado na seção 2.5.. Embora a metodologia de simulação proposta em seu artigo seja diferente, os

resultados encontrados se assemelham, pois em ambos os trabalhos, a distribuição das estimativas obtidas ficou centrada no verdadeiro valor populacional. Além disso, ele comparou as estimativas obtidas pelo método RDS com as estimativas se tivesse obtido a amostra utilizando amostragem aleatória simples, e a variabilidade para as estimativas RDS também foi maior do que a variabilidade para a amostragem aleatória simples. Embora nessa dissertação, todas as amostras tenham sido obtidas utilizando a metodologia RDS, esses resultados também foram verificados, com as estimativas obtidas de forma simples apresentando variabilidade menor do que as estimativas denominadas RDS. O outro estudo (Salganik & Heckathorn, 2004), que assim como nessa dissertação, usou a abordagem *model-driven*, também teve seu algoritmo apresentado na seção 2.5.. Um de seus resultados mostrou que quanto maior o tamanho da amostra, mais próximas do verdadeiro valor do parâmetro ficaram as estimativas. No entanto, ao contrário do apresentado aqui, para amostras menores, sempre havia uma tendência de superestimar a prevalência. As figuras 5.8., 5.9., 5.10. e 5.11. mostraram que os valores podem apresentar-se superestimados ou subestimados, independente do tamanho amostral. Vale destacar ainda que mesmo nos casos de superestimação apresentados por Salganik & Heckathorn (2004), a diferença entre a estimativa e o verdadeiro valor do parâmetro eram muito pequenas, da ordem da terceira casa decimal, diferente do que aconteceu com dados apresentado no capítulo 5. Uma das razões para isso pode ser o fato do trabalho citado utilizar amostragem com reposição, à medida que as simulações apresentadas nessa dissertação foram geradas por processos sem reposição.

Finalmente, as considerações sobre o método de ponderação para as estimativas. Para avaliar esse método, é importante considerar cada tipo de infecção. Como foi apresentado no capítulo 5, para infecções onde o tamanho das redes de contato têm influência na exposição à doença, e consequentemente nas chances de se contrair a doença (infecção B), a ponderação proposta pelo modelo Heckathorn apresentou ótimos resultados, pois o modelo simples apresentou uma tendência em superestimar as estimativas. Além disso, a ponderação também permitiu obter resultados mais precisos nas situações onde a infecção foi criada por contágio (infecção D). Ambos os casos citados acima se encaixam bem nas situações usualmente estudadas com populações ocultas, pois imagina-se que de alguma forma, pessoas que conhecem mais gente, estão mais vulneráveis a contrair uma infecção, assim como ao se estudar uma doença sexualmente transmissível, é necessário que haja o contato entre as pessoas para o

espalhamento da doença. Nesse sentido, ter um bom acesso a essas populações é muito importante, principalmente porque isso facilita a correta estimação de prevalência dessas doenças, o que viabiliza, por exemplo, o conhecimento do perfil dessas populações e a realização de intervenções mais efetivas. Assim, tem-se que a ponderação proposta para estimar prevalências contribui para estimativas mais fidedignas, pelo menos para características dicotômicas, como é o caso de prevalências. Estimadores que abordam outras propostas de ponderação também estão sendo desenvolvidos (Heckathorn, 2007 e Volz & Heckathorn, 2008). Assim, realizar simulações para testá-los também é uma idéia a ser desenvolvida futuramente.

Por outro lado, embora as conclusões desse trabalho estejam sendo bastante favoráveis ao uso do RDS, é necessário ainda citar algumas outras observações feitas por Gile & Handcock (2009). Em todas as simulações realizadas por eles, haviam três formas de seleção das sementes: todas eram infectadas, nenhuma era infectada ou as sementes eram escolhidas aleatoriamente, independente da infecção. Além de verificar o comportamento das estimativas de prevalência considerando o número de ondas da amostra, os autores também simularam situações onde o grau de interconectividade, ou seja, a homofilia, era baixo ou alto. Em geral, para amostras que partiram de sementes não infectadas, as estimativas de prevalência apresentaram-se subestimadas. Da mesma forma, para amostras com sementes infectadas, os resultados apresentaram-se superestimados. Finalmente, para amostras onde as sementes foram selecionadas aleatoriamente, a distribuição das estimativas de prevalência estava centrada no verdadeiro valor populacional de 0,2. Vale destacar ainda que para essas simulações, foi utilizado o estimador proposto por Volz & Heckathorn (2008), e que, ao comparar as estimativas utilizando esse estimador e o estimador utilizado nessa dissertação, os autores verificaram maior eficiência para o estimador mais atual. Com isso, mais uma possibilidade que pode ser desenvolvida futuramente é a reprodução das simulações realizadas por eles, focando nos mesmos parâmetros de observação, a fim de buscar resultados comparáveis nessas duas abordagens.

Adicionalmente, Gile & Handcock (2009) observaram a precisão das estimativas relacionando-as com o tamanho verdadeiro da população e verificaram uma menor variabilidade das estimativas, quanto maior o percentual da amostra em relação à população, considerando que tanto as pessoas infectadas, como as não infectadas, possuem a mesma média de contatos. Por outro lado, a medida que as pessoas infectadas possuem médias de contato maiores do que aquelas não infectadas, foi

verificada uma subestimação das prevalências. Essa questão também não foi abordada nessa dissertação, já que as populações criadas tinham o mesmo tamanho (25.000 indivíduos) e as amostras, aproximadamente também.

Embora a metodologia RDS esteja sendo bastante utilizada na estimação de prevalências de HIV e outras DST's, na literatura ainda existem poucos trabalhos que abordam a eficiência das estimativas geradas, ou seja, o quão precisas e verdadeiras elas são. Assim, essa dissertação abordou alguns aspectos sobre essa eficiência, contribuindo para a afirmação de que a metodologia é válida e pode produzir bons resultados, ainda que deva ser aplicada com alguma cautela.

No entanto, as conclusões dessa dissertação ficam um pouco limitadas, pois como já mencionado, cada população oculta apresenta características bastante distintas e aqui, as simulações realizadas utilizaram dados de apenas um estudo e uma população alvo – homens que fazem sexo com homens (HSH). Além disso, embora alguns trabalhos conjugando técnicas de simulação e RDS já estejam disponíveis, resultados que abordem a mesma metodologia dessa dissertação, ou seja, considerando estruturas populacionais mais complexas, não foram encontrados na literatura. Nesse sentido, outra proposta para trabalhos futuros é a repetição da metodologia aqui desenvolvida, que pode ser aplicada na investigação de outras populações ocultas, e em seguida, pode ser feita também uma comparação entre os resultados obtidos para cada população.

7. Referências bibliográficas

Anderson R. (1996) *The spread of HIV and sexual mixing patterns*. Aids in the world II. Capítulo 4. Editado por Jonathan Mann e Daniel Tarantola.

Bailey, N. (1958) *The mathematical theory of epidemics*. Biometrika, 1958: 45(3-4):589

Barabási A, Albert R. (1999) *Emergence of scaling in random networks*. Science, 286:509-512

Bollobás B. (2001) *Random graphs*. Editora Cambridge University. 2ª edição. Disponível no site http://books.google.com/books?id=o9WecWgilzYC&hl=pt-BR. (acessado em 06 de junho de 2009)

Borgatti S. (2009) *A brief guide to using NetDraw*. Disponível no site http://www.analytictech.com/Netdraw/netdraw.htm (último acesso em 14 de junho de 2009).

Bussab W, Morettin P. (2007) *Estatística Básica*. Edtora Saraiva, 5a edição. São Paulo. Página 256.

Díaz A, Barruti M, Doncel C (1992). *The Lines of Success? A study on the nature and extent of cocaine use in Barcelona*. Barcelona: Laboratori de Sociologia (ICESB).

Friedman SR, Bolyard M *et al* (2007). *Some data-driven reflections on priorities in aids network research.* AIDS Behav 2007; 11: 641-651.

Gile K, Handcock M. (2009) *Respondent-driven sampling: an assessment of current methodology*. E-print disponível no site http://arxiv.org/abs/0904.1855 (acessado em 14 de junho de 2009).

Goodman L. (1961) *Snowball sampling*. Annals of Mathematical Statistics, 32: 148-170

Hartmann S. (2005) *The world as a process: Simulations in the natural and social sciences*. PhilSci Archive. Artigo disponível no site http://philsci-archive.pitt.edu/archive/00002412/ (acessado em 13 de junho de 2009)

Heckatorn D. (1997) *Respondent-driven sampling: a new approach to the study of hidden populations*. Social Problems, 1997; 44:174-199

_____ (2002) *Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hideen populations*. Social Problems, 2002; 49:11-34.

_____ (2007) *Extensions of Respondent-driven Sampling: analyzing continuous variables and controlling for differential recruitment using dual-component sampling weights*. Sociological Methodology, 37: 151-207

Heckathorn D, Semaan S, Broadhead R, Hughes J. (2002) *Extensions of Respondent-Driven sampling: a new approach to the study of injection drug users aged 18-25*. Aids and Behavior 2002, (6)1:55-67.

Keeling M, Eames K. (2005) *Networks and epidemic models*. J. R. Soc. Interface (2005) 2, 295-307.

Kemeny J, Snell J. (1960) *Finite Markov chains*. Princeton, N.J.: Van Nostrand.

Killworth P, Bernard H. (1978/79) *The reversal small-world experiment*. Social Networks, 1 (1978/79) 159-192

Klovdhal AS. (1985) *Social networks and the spread of infectious diseases: the AIDS example.* Soc Sci Med 1985, 21: 1203-1206

Koopman J. (2004) *Modeling infection transmission*. Annu Rev Public Health 2004; 25:303-326

Laird S, Jensen H. (2006) *A non-growth network model exponential and 1/k scale-free degree distribution*. Europhys. Lett., 76(4): 710-716

Luke DA, Harris JK. (2007) *Network analysis in Public Health: history, methods and applications.* Annu Rev Public Health 2007; 28:69-93.

Magnani R, Sabin K, Saidel T, Heckathorn D. (2005) *Review of sampling hard-to-reach hidden populations for HIV surveillance*. AIDS 2005, 19 (supll 2): S67-S72

Mello M, Pinho A, Chinaglia M, Tun W, Barbosa Júnior A, Ilário M, Reis P, Salles R, Westman S, Díaz R. (2008) *Assessment of risk factors for HIV infection among men who have sex with men in the metropolitan area of Campinas city, Brazil, using Respondent-Driven Sampling*. Relatório técnico disponível no site http://www.popcouncil.org/horizons/projects/Brazil_MSMRiskFactors.htm (acessado em 06 de junho de 2009)

Meyers L, Newman M, Martin M, Schrag S. (2003) *Applying network theory to epidemics: Control measures for Mycoplasma peneumoniae outbreaks*. Emerging Infectious Diseases 9, 204-210

Meyers L, Newman M, Pourbohloul, B. (2006) *Predicting epidemics on directed contact networks*. Journal of Theoretical Biology, 240 (2006) 400-418

Morris M. (2004) *Network Epidemiology: A handbook for survey design and data collection.* Editora Okford University. Inglaterra, páginas 8-21.

Newman M, Strogatz S, Watts D. (2001) *Random graphs with arbitrary degree distributions and their applications*. Physical Review E, 64, 026118

Newman M, Watts D, Strogatz S. (2002) *Random graph models of social networks*. Proceedings of the National Academy of Sciences of the United States of America. Vol 99, suppl 1: 2566-2572

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramirez-Valles J, Heckathorn D, Vásquez R, Diaz R, Campbell R. (2005) *From networks to populations: the development and application of Respondent-Driven Sampling among IDU's and latino gay men*. Aids and Behavior 2005, (9)4:387-402.

Robinson W, Risser J, McGoy S, Becker A, Rehman H, Jefferson M, Griffin V, Wolverton M, Tortu S. (2006) *Recruiting injection drug users: a three-site comparison of results and experiences with respondent-driven and target sampling procedures*. Journal of Urban Health: Bulletin of the New York Academy of Medicine.

Ross, R. (1916) *An application of the theory of probabilities to the study of a priori pathometry, II* Proc R Soc 1916; A92: 204-230.

Rothman, K and Greenland, S (1998). *Modern Epidemiology*. Editora Lippincott – Raven, 2ª edição, Filadélfia, página 30.

Salganik, M (2006) *Variance estimation, design effects, and sample size calculations for respondent-driven sampling*. Journal of Urban Health 2006; 83 (6 Suppl):i98-112

Salganik, M and Heckathorn D. (2004) *Sampling and estimation in hidden populations using respondent-driven sampling*. Sociological Methodology 34: 193-239

Semaan S, Lauby J, Liebman J. (2002) *Street and network sampling in evaluation studies of HIV risk-reduction interventions*. Aids Rev 2002; 4:213-223

Scott J. (2000) *Social networks analysis*. Editora SAGE Publications Ltd., 2ª edição, Reino Unido, páginas 8-10.

Simões A, Bastos F, Moreira R, Lynch K, Metzger D. (2006) *A randomized trial of audio computer and in-person interview to assess HIV risk among drug and alcohol users in Rio de Janeiro, Brazil.* Journal of Substance Abuse Treatment, 30 (2006): 237-243

Singer M, Stopka T, Siano C, Springer K, Barton G, Khoshnood K, Gorry de Puga A, Heimer R. (2000) *The social geography of AIDS anda hepatitis risk: qualitative approaches for assessing local differences in sterile-syringe access among injection drug users*. American Journal of Public Health. 2000;90:1049-1056

Strogatz S. (2001) *Exploring complex networks*. Nature Vol. 410: 268-276

Stumpf M, Wiuf C. (2005) *Sampling properties of random graphs: the degree distribution*. Physical review, E 72, 036118 (2005)

Volz E. (2004) *Random networks with tunable degree distribution and clustering*. Physical Review E, 70, 056115.

Volz E, Heckathorn D. (2008) *Probability based estimation theory for Respondent-driven sampling*. Journal of Official Statistics, Vol. 24, No. 1, 2008, 79-97

Volz E, Werjnert C, Degani I, Heckathorn D. (2007) *Respondent-Driven Sampling analysis tool (RDSat)*. Versão 5.6.

Wallinga J, Edmunds WJ, Kretzschmar M (1999). *Perspective: human contact patterns and the spread of airbone infectious diseases*. Trends in Microbiology 1999; 7(9): 372-377.

Wattana W, Griensven F, Rhucharoenpornpanich O, Manopaiboon C, Thienkrua W, Bannatham R, Fox K, Mock P, Tappero J, Levine W. (2007) *Respondent-driven sampling to assess characteristics and estimate the number os injection drug users in Bangkok, Thailand*. Drug and Alcohol Dependence, 90(2007) 228-233.

Watts, D. (2003) *Six degrees. The science of a connected age*. Editora W.W. Norton. 1ª edição, New York. Capítulo 1, páginas 27-29.

Watts D. (2004) *Small worlds.* Editora Princeton University. 1ª edição. Princeton, New Jersey.

Wylie J, Jolly A. (2001) *Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada*. Sex Transm Dis 2001; 28:14-24

ANEXO I. Scripts utilizados para a geração das populações e casos infectados.

(A) Construção da população inicial de 50.000 elementos.

```
# Carrega dados originais (base para os parâmetros)
load('dissertacao.RData')

### Construcao da populacao
## Selecionar aleatoriamente algumas caracteristicas e parear apenas
classe economica e escolaridade - Essas duas variaveis sao usadas
apenas na construcao da rede de infeccao.

# Cria o banco para a população
N = 25000
pop <- data.frame(cbind(id=1:N, idade=NA, orient=NA, grau=NA,
clecon=NA, escol=NA))

# grau dos individuos
lambda = 0.08
pop$grau<-round(rexp(N,lambda))+1

# idade
pop$idade <- sample(dados$age.part,N,replace=TRUE)

# orientacao sexual
pbi = prop.table(table(dados$sexorient.part[dados$sexorient.part==1 |
dados$sexorient.part==3]))[2]
pop$orient[1:round(pbi*N)]<-'bi'
pop$orient[(round(pbi*N)+1):N]<-'homo'

# classe economica e escolaridade
pop.temp <- dados[sample(1:dim(dados)[1],N,replace=TRUE),]
pop$clecon <- pop.temp$clecon.part
pop$escol <- pop.temp$escol.part
for (i in 1:dim(pop)[1]) {if (is.na(pop$clecon[i]==TRUE))
pop$clecon[i] <-0}
for (i in 1:dim(pop)[1]) {if (is.na(pop$escol[i]==TRUE)) pop$escol[i]
<-0}

# Salvar esse arquivo
library(foreign)
write.dbf(pop, "populacao.dbf")
save.image("C:\\populacao_final.RData")
```

(B) Construção da estrutura de Ligação 01 (rede aleatória).

```
### 01 - Carrega a populacao
load('populacao_final.RData')

# Cria as ligações dos pares da população de forma aleatória
id1 <- data.frame(id=rep(1:N,times=c(pop$grau[1:N])),amigo=NA)
```

```
ligacoes <- data.frame(id=id1[,1],peso=NA)
k <- 1
escolhidos<-0
fim <- c(0,0)

for (j in 1:(N-1)){
  lig.temp <- ligacoes[ligacoes$id ==j,]
  ligacoes <- ligacoes[ligacoes$id !=j,]
  if (dim(lig.temp)[1]!=0){
   for (i in 1:dim(lig.temp)[1]) {
     pos = sample(1:dim(ligacoes)[1],1)
     a = ligacoes[pos,1]
     tentativa = 0
     while (sum(a == escolhidos)!=0 & tentativa < N/2) {
      pos = sample(1:dim(ligacoes)[1],1)
      a = ligacoes[pos,1]
      tentativa = tentativa + 1
     }
     escolhidos <- c(escolhidos,a)
     ligacoes <- ligacoes[-pos,]
     fim <- rbind(fim, c(j,a),c(a,j))
   }
  escolhidos <- 0
  print(j)
  }
}

# Cria o banco das ligações
fim<-fim[-1,]
id2 <- data.frame(id=fim[,1], amigo=fim[,2])
```

(C) Construção da estrutura de Ligação 02 (rede baseada na orientação sexual).

```
### 01 - Carrega a populacao
load('populacao_final.RData')
load('dissertacao.RData')

# Cria as ligações dos pares da população com base na orientação
sexual
bi=table(pop$orient)[1]
id1 <- data.frame(id=rep(1:N,times=c(pop$grau[1:N])),
         orient=c(rep(2,times=sum(pop$grau[1:bi])),rep(4,times=
sum(pop$grau[(bi+1):N]))),
         amigo=NA)
         # orient = 2 é 'bi' e orient= 4 é 'homo'

# Probabilidades de escolha
orientsex <- dados[(dados$sexorient.part==1 |
dados$sexorient.part==3) & (dados$sexorient.conv==1 |
dados$sexorient.conv==3),]
table(orientsex$sexorient.part, orientsex$sexorient.conv)
prop.table(table(orientsex$sexorient.part, orientsex$sexorient.conv),
margin=2)

# Homo escolher homo
```

```
p1 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[1]
#Homo escolher bi
p2 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[2]
# Bi escolher homo
p3 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[3]
#Bi escolher bi
p4 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[4]

id1$pbiconv <- ifelse(id1$orient==4,id1$pbiconv<-p3, id1$pbiconv<-p4)
table(id1$pbiconv)
id1$phomoconv <- ifelse(id1$orient==4,id1$phomoconv<-
p1,id1$phomoconv<-p2)
table(id1$phomoconv)
head(id1)

ligacoes <- id1[,c(1,2,4,5)]
k <- 1
escolhidos<-0
fim <- c(0,0)

for (j in 1:(N-1)){
  lig.temp <- ligacoes[ligacoes$id ==j,]
  ligacoes <- ligacoes[ligacoes$id !=j,]
  if (dim(lig.temp)[1]!=0){

   for (i in 1:dim(lig.temp)[1]) {
     if (lig.temp$orient[i]==2) peso<- ligacoes$pbiconv else peso<-
ligacoes$phomoconv
     pos = sample(1:dim(ligacoes)[1],1,prob=peso)
     a = ligacoes[pos,1]
     tentativa = 0
     while (sum(a == escolhidos)!=0 & tentativa < N/2) {
      pos = sample(1:dim(ligacoes)[1],1,prob=peso)
      a = ligacoes[pos,1]
      tentativa = tentativa + 1
     }

     escolhidos <- c(escolhidos,a)
     ligacoes <- ligacoes[-pos,]
     fim <- rbind(fim, c(j,a),c(a,j))
   }
  escolhidos <- 0
  }
}

# Cria o banco das ligações
fim<-fim[-1,]
id2 <- data.frame(id=fim[,1], amigo=fim[,2])
```

(D) Construção da estrutura de Ligação 03 (rede baseada na idade).

```
### 01 - Carrega a populacao
```

```
load('populacao_final.RData')

## ligaçõoes com base na distribuição de idades
id1 <- data.frame(id=rep(1:N,times=c(pop$grau[1:N])),
         idade=rep(c(pop$idade),times=c(pop$grau[1:N])),
         amigo=NA)

ligacoes <- data.frame(id=id1[,1],peso=NA, idade=id1[,2])
k <- 1
escolhidos<-0
fim <- c(0,0)


for (j in 1:(N-1)){
  lig.temp <- ligacoes[ligacoes$id ==j,]
  ligacoes <- ligacoes[ligacoes$id !=j,]

  if (dim(lig.temp)[1]!=0){

   for (i in 1:dim(lig.temp)[1]) {
     peso <- c(dnorm(c(ligacoes$idade),
mean=(13.57+0.45*(lig.temp$idade[i])), sd=sqrt(47.3)))
     pos = sample(1:dim(ligacoes)[1],1,prob=peso)
     a = ligacoes[pos,1]
     tentativa = 0
     while (sum(a == escolhidos)!=0 & tentativa < N/2) {
      pos = sample(1:dim(ligacoes)[1],1, prob=peso)
      a = ligacoes[pos,1]
      tentativa = tentativa + 1
     }

     escolhidos <- c(escolhidos,a)
     ligacoes <- ligacoes[-pos,]
     fim <- rbind(fim, c(j,a),c(a,j))
   }
  escolhidos <- 0
  print(j)
  }
}

# Cria o banco das ligações
fim<-fim[-1,]
id2 <- data.frame(id=fim[,1], amigo=fim[,2])
```

(E) Construção da estrutura de Ligação 04 (rede baseada em sexo e idade).

```
### 01 - Carrega a populacao
load('populacao_final.RData')
load('dissertacao.RData')

# Cria as ligações dos pares da população considerando a distribuicao
etaria e a orientacao sexual
bi=table(pop$orient)[1]
id1 <- data.frame(id=rep(1:N,times=c(pop$grau[1:N])),
         orient=c(rep(2,times=sum(pop$grau[1:bi])),rep(4,times=
sum(pop$grau[(bi+1):N]))),
```

```
          idade=rep(c(pop$idade),times=c(pop$grau[1:N])),
          amigo=NA)
          # orient = 2 ï¿½ 'bi' e orient= 4 ï¿½ 'homo'

# Probabilidades de escolha
orientsex <- dados[(dados$sexorient.part==1 |
dados$sexorient.part==3) & (dados$sexorient.conv==1 |
dados$sexorient.conv==3),]
table(orientsex$sexorient.part, orientsex$sexorient.conv)
prop.table(table(orientsex$sexorient.part, orientsex$sexorient.conv),
margin=2)

# Homo escolher homo
p1 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[1]
#Homo escolher bi
p2 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[2]
# Bi escolher homo
p3 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[3]
#Bi escolher bi
p4 <- prop.table(table(orientsex$sexorient.part,
orientsex$sexorient.conv), margin=2)[4]

id1$pbiconv <- ifelse(id1$orient==4,id1$pbiconv<-p3, id1$pbiconv<-p4)
table(id1$pbiconv)
id1$phomoconv <- ifelse(id1$orient==4,id1$phomoconv<-
p1,id1$phomoconv<-p2)
table(id1$phomoconv)
head(id1)

ligacoes <- id1[,c(1,2,3,5,6)]
k <- 1
escolhidos<-0
fim <- c(0,0)

for (j in 1:(N-1)){
  lig.temp <- ligacoes[ligacoes$id ==j,]
  ligacoes <- ligacoes[ligacoes$id !=j,]
  if (dim(lig.temp)[1]!=0){

   for (i in 1:dim(lig.temp)[1]) {
     if (lig.temp$orient[i]==2) peso<- ligacoes$pbiconv else peso<-
ligacoes$phomoconv
     idade <- c(dnorm(c(ligacoes$idade),
mean=(13.57+0.45*(lig.temp$idade[i])), sd=sqrt(47.3)))
     peso <- peso*idade

     pos = sample(1:dim(ligacoes)[1],1,prob=peso)
     a = ligacoes[pos,1]
     tentativa = 0
     while (sum(a == escolhidos)!=0 & tentativa < N/2) {
      pos = sample(1:dim(ligacoes)[1],1,prob=peso)
      a = ligacoes[pos,1]
      tentativa = tentativa + 1
     }

     escolhidos <- c(escolhidos,a)
     ligacoes <- ligacoes[-pos,]
     fim <- rbind(fim, c(j,a),c(a,j))
```

```
   }
  escolhidos <- 0
  print(j)
  }
}


# Cria o banco das ligações
fim<-fim[-1,]
id2 <- data.frame(id=fim[,1], amigo=fim[,2])
```

(F) Determinação dos casos infectados – Infecção A (aleatória simples).

```
### 02 - Selecionar os indivíduos infectados na população
#Prevalencia
p = 0.2
casos = round(N*p)
am <- sort(sample(1:dim(pop)[1],casos,replace=FALSE))
pop$infec1=0
pop$infec1[am]=1
```

(G) Determinação dos casos infectados – Infecção B (aleatória ponderada, com probabilidade de seleção proporcional ao grau).

```
### 02 - Selecionar os indivíduos infectados na população
#Prevalencia
p = 0.2
casos = round(N*p)

am <- sort(sample(1:dim(pop)[1],casos,replace=FALSE, prob=pop$grau))
pop$infec2=0
pop$infec2[am]=1
```

(H) Determinação dos casos infectados – Infecção C (aleatória ponderada, com probabilidade de seleção determinada por covariáveis).

```
### 02 - Selecionar os indivíduos infectados na população
#Prevalencia
p = 0.2
casos = round(N*p)

# Obtenção dos parametros para a seleção dos indivíduos
# m2 <- glm(hiv2.part~ age.part + as.factor(clecon.part) +
escol.part, family="binomial", data=dados)
# summary(m2)

attach(pop)
pop$peso <- ifelse(clecon==1, exp(idade*0.076-
escol*0.11091)/(1+exp(idade*0.076-escol*0.11091)) ,
       ifelse(clecon==2, exp(idade*0.076-escol*0.11091-
clecon*1.1952)/(1+exp(idade*0.076-escol*0.11091-clecon*1.1952)) ,
       ifelse(clecon==3, exp(idade*0.076-escol*0.11091-
clecon*1.3989)/(1+exp(idade*0.076-escol*0.11091-clecon*1.3989)),
       exp(idade*0.076-escol*0.11091)/(1+exp(idade*0.076-
escol*0.11091)) )))
```

```
detach(pop)

am <- sort(sample(1:dim(pop)[1],casos,replace=FALSE, prob=pop$peso))
pop$infec3=0
pop$infec3[am]=1
```

(I) Determinação dos casos infectados – Infecção D (cluster).

```
### 02 - Selecionar os indivíduos infectados na população
#Prevalencia
pop$id_original <- pop$id
pop$id <- 1:N
p = 0.2
casos = round(N*p)

# Tempo 0 = Todos sao suscetiveis
pop$infec4 <- 0
id2$infec <- 0

# Tempo 1 - Escolhe-se aleatoriamente algumas pessoas para se
infectarem
n1 = 50
amostra_t1 <- sort(sample(1:N,n1,replace=FALSE))
pop$infec4[amostra_t1]<-1

for (i in amostra_t1) id2$infec[id2$id==i]<-1
#table(id2$infec,id2$id)

# Nos tempos seguintes - Escolhe-se segundo uma probabilidade os
amigos que serao infectados
beta = 0.7

while (sum(pop$infec4)<casos){
# calculando a prob de infectar
a <- aggregate(id2$infec,by=list(id2$amigo),sum)
a <- data.frame(amigo=a[,1],infe2=a[,2],grau=pop$grau)
a$prob<-beta*a$infe2/a$grau

moeda = runif(N)
a$novoscasos <- as.numeric(moeda<a$prob)
pop$infec4[a$novoscasos==1]<-1
for (i in a$amigo[a$novoscasos==1]) id2$infec[id2$id==i]<-1
print(sum(pop$infec4))
}

sorteio <- sort(a$amigo[a$novoscasos==1])
b <- sort(sample(sorteio, (sum(pop$infec4)-casos)))
for (i in 1:length(b)) pop$infec4[pop$id==b[i]] <- 0
```

ANEXO II. Scripts utilizados para a implementação do processo de amostragem.

(A) Recrutamento completo.

```
### 03 - Amostragem (recrutamento completo)
# Definindo o tamanho mínimo da amostra
n <- 500

# Determinando o número de sementes
s = 5

# Escolhendo as sementes
sem = sample(pop$id[pop$grau>10],s,prob=pop$grau[pop$grau>10])
# Escolhendo os convites entregues por cada semente
convi <- sample(0:3,s,prob=c(0.47,0.24,0.18,0.11), replace=TRUE)
# Identificando as sementes como RDS
id.rds <- data.frame(partic=sem,id=1:s,onda=0,i_rec=0, i_part=0)


# ONDA 1
amostra_final <- c(sem)
onda1 <- NULL

for (i in 1:s){
 id2$am <- 0
 id.rds$i_part[i] <- pop$infec1[pop$id==sem[i]]

 for (j in 1:length(amostra_final))
id2$am[id2$amigo==amostra_final[j]]<-1
 grupo <- id2$amigo[(id2$id==sem[i] & id2$am !=1)]
 print(length(grupo))
 if ((length(grupo) < convi[i]) & (length(grupo) != 0) & (convi[i]
!=0))
  {am_grupo <- grupo
   id.rds.temp <- data.frame(partic=am_grupo,
id=i*10+c(1:length(am_grupo)), onda=1, i_rec=0, i_part=0)
   for (k in 1:length(am_grupo))
   {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
    id.rds.temp$i_rec[k] <- pop$infec1[pop$id==sem[i]]}
   id.rds <- rbind(id.rds, id.rds.temp)
   }
  else
   if(convi[i]!=0 & length(grupo)!=0) {am_grupo <-
sample(grupo,convi[i], replace=FALSE)
   id.rds.temp <- data.frame(partic=am_grupo, id=i*10+c(1:convi[i]),
onda=1, i_rec=0, i_part=0)
   for (k in 1:length(am_grupo))
   {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
    id.rds.temp$i_rec[k] <- pop$infec1[pop$id==sem[i]]}
   id.rds <- rbind(id.rds, id.rds.temp)}
  if (convi[i]!=0 & length(grupo)!=0)
   {amostra_final <- c(amostra_final, am_grupo)
    onda1 <- c(onda1, am_grupo)}
}

# A partir da ONDA 2, usar essa rotina:
on <- 2
```

```
while ((length(amostra_final) < n) & (sum(convi) !=0))
{

b <- id.rds$partic[id.rds$onda==(on-1)]
if (length(b)!=0){
  # Escolhendo o número de convites entregues por cada participante
  convi <- sample(0:3,length(b),prob=c(0.47,0.24,0.18,0.11),
replace=TRUE)

  for (i in 1:length(b)){

   if (convi[i] !=0){
    id2$am <- 0
    for (j in 1:length(amostra_final))
id2$am[id2$amigo==amostra_final[j]]<-1
    grupo <- id2$amigo[(id2$id==b[i] & id2$am !=1)]
    print(length(grupo))
    if ((length(grupo) < convi[i]) & (length(grupo) != 0))
     {
                                              am_grupo <- grupo
     id.rds.temp <- data.frame(partic=am_grupo,
id=id.rds$id[id.rds$onda==(on-1)][i]*10+c(1:length(am_grupo)),
onda=on, i_rec=0, i_part=0)
     for (k in 1:length(am_grupo))
      {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
      id.rds.temp$i_rec[k] <- id.rds$i_part[id.rds$partic==b[i]]
                                                } # end for
     id.rds <- rbind(id.rds, id.rds.temp)
     } # end if
    else
    if(length(grupo)!=0)
      {am_grupo <- sample(grupo,convi[i], replace=FALSE)
      id.rds.temp <- data.frame(partic=am_grupo,
id=id.rds$id[id.rds$onda==(on-1)][i]*10+c(1:convi[i]), onda=on,
i_rec=0, i_part=0)
      for (k in 1:length(am_grupo))
       {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
       id.rds.temp$i_rec[k] <- pop$infec1[pop$id==b[i]]
                                                 } # end for
      id.rds <- rbind(id.rds, id.rds.temp)
                                                } # end if
    if (length(grupo)!=0) amostra_final <- c(amostra_final, am_grupo)
   } # end if conv[i]
  } # end for i
  on <- on+1
  }
if (length(b)==0){
  convi=0 # truque para interromper a amostragem quando nao houver
mais ninguem para convidar
  print ('WARNING: a cadeia de referencia terminou antes de atingir n
na replica:')
  print(q)
  }
}

# Cria a variável que representa quem entrou na amostra
pop$amostraRDS <- 0
pop$amostraRDS[sort(amostra_final)] <- 1
```

(B) Recrutamento aleatorizado.

```
### 03 - Amostragem (todos entregam todos os convites)
# Definindo o tamanho mínimo da amostra
n <- 500

# Determinando o número de sementes
s = 5

# Escolhendo as sementes
sem = sample(pop$id[pop$grau>10],s,prob=pop$grau[pop$grau>10])
# Identificando as sementes como RDS
id.rds <- data.frame(partic=sem,id=1:s,onda=0,i_rec=0, i_part=0)


# ONDA 1
amostra_final <- c(sem)
onda1 <- NULL

for (i in 1:s){
 id2$am <- 0
 id.rds$i_part[i] <- pop$infec1[pop$id==sem[i]]

 for (j in 1:length(amostra_final))
id2$am[id2$amigo==amostra_final[j]]<-1
 grupo <- id2$amigo[(id2$id==sem[i] & id2$am !=1)]
 print(length(grupo))
 if ((length(grupo) < 3) & (length(grupo) != 0))
  {am_grupo <- grupo
   id.rds.temp <- data.frame(partic=am_grupo,
id=i*10+c(1:length(am_grupo)), onda=1, i_rec=0, i_part=0)
   for (k in 1:length(am_grupo))
   {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
    id.rds.temp$i_rec[k] <- pop$infec1[pop$id==sem[i]]}
   id.rds <- rbind(id.rds, id.rds.temp)
   }
  else
   if(length(grupo) !=0) {am_grupo <- sample(grupo,3, replace=FALSE)
   id.rds.temp <- data.frame(partic=am_grupo, id=i*10+c(1:3), onda=1,
i_rec=0, i_part=0)
   for (k in 1:length(am_grupo))
   {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
    id.rds.temp$i_rec[k] <- pop$infec1[pop$id==sem[i]]}
   id.rds <- rbind(id.rds, id.rds.temp)}
  amostra_final <- c(amostra_final, am_grupo)
  onda1 <- c(onda1, am_grupo)
}

# A partir da ONDA 2, usar essa rotina:
on <- 2
on_max <- 20

while ((length(amostra_final) < n) & (on < on_max) )
{

b <- id.rds$partic[id.rds$onda==(on-1)]

for (i in 1:length(b)){

 id2$am <- 0
```

```
 for (j in 1:length(amostra_final))
id2$am[id2$amigo==amostra_final[j]]<-1
 grupo <- id2$amigo[(id2$id==b[i] & id2$am !=1)]
 print(length(grupo))
 if ((length(grupo) < 3) & (length(grupo) != 0))
  {am_grupo <- grupo
   id.rds.temp <- data.frame(partic=am_grupo,
id=id.rds$id[id.rds$onda==(on-1)][i]*10+c(1:length(am_grupo)),
onda=on, i_rec=0, i_part=0)
   for (k in 1:length(am_grupo))
   {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
    id.rds.temp$i_rec[k] <- id.rds$i_part[id.rds$partic==b[i]]}
   id.rds <- rbind(id.rds, id.rds.temp)
  }
  else
   if(length(grupo) !=0) {am_grupo <- sample(grupo,3, replace=FALSE)
   id.rds.temp <- data.frame(partic=am_grupo,
id=id.rds$id[id.rds$onda==(on-1)][i]*10+c(1:length(am_grupo)),
onda=on, i_rec=0, i_part=0)
    for (k in 1:length(am_grupo))
    {id.rds.temp$i_part[k] <- pop$infec1[pop$id==am_grupo[k]]
     id.rds.temp$i_rec[k] <- pop$infec1[pop$id==b[i]]}
   id.rds <- rbind(id.rds, id.rds.temp)}
  amostra_final <- c(amostra_final, am_grupo)
  }
on <- on+1
}

# Cria a variável que representa quem entrou na amostra
pop$amostraRDS <- 0
pop$amostraRDS[sort(amostra_final)] <- 1
```

ANEXO III. Script utilizado para a obtenção das estimativas de prevalência nas amostras.

```
### 04 - Estimar a prevalência na amostra desconsiderando o desenho
do estudo (estimativa simples)

banco_est <- pop[pop$amostraRDS==1,]
estimativa <- sum(banco_est$infec1)/dim(banco_est)[1]
estimativa


### 05 - Estimar a prevalência na amostra considerando o desenho do
estudo (estimativa RDS)

# Numero de pessoas em cada grupo
na <- sum(banco_est$infec1)
nb <- dim(banco_est)[1]- na
#Media dos graus das pessoas em cada grupo
banco_est$peso <- 1/(banco_est$grau)
peso_na <- sum(banco_est$peso[banco_est$infec1==1])
peso_nb <- sum(banco_est$peso[banco_est$infec1==0])
Da <- na/peso_na
Db <- nb/peso_nb

# ligacoes: 1 - AcomA ; 2 - BcomB ; 3 - AcomB
id.rds$rel <- ifelse((id.rds$i_rec==1 &
id.rds$i_part==1),id.rds$rel<-1,
      ifelse((id.rds$i_rec==0 & id.rds$i_part==0),id.rds$rel <- 2,
      ifelse((id.rds$i_rec==1 & id.rds$i_part==0),id.rds$rel <-3,
      ifelse((id.rds$i_rec==0 & id.rds$i_part==1),id.rds$rel <-
4,5))))
id.rds$um <- 1
liga <- id.rds[-c(1:s),]
raa <- sum(liga$um[liga$rel==1])
rab <- sum(liga$um[liga$rel==3])
rba <- sum(liga$um[liga$rel==4])
rbb <- sum(liga$um[liga$rel==2])
Cab <- rab/(rab+raa)
Cba <- rba/(rba+rbb)

# Estimativa
PPa <- (Db*Cba) / (Db*Cba+Da*Cab)
PPa
```