Pensamiento Psicológico, Vol. 4, N°10, 2008, pp. 167-176

A comparison of three methods for sampling hard - to - reach or hidden populations

Bernardo Useche Aldana[1]
University of Texas, Health Science Center- Houston (Estados Unidos)

Marcela Arrivillaga Quintero[2]
Pontificia Universidad Javeriana-Cali (Colombia)

Recibido: 07/12/07 Aceptado: 14/02/08

Resumen

Los proyectos de investigación que necesitan reclutar participantes pertenecientes a poblaciones "ocultas" o "difíciles de encontrar" requieren de métodos de muestreo que no solo faciliten la recolección de los datos y la confidencialidad, sino también que incrementen la validez externa y permitan hacer inferencias estadísticas apropiadas. Este artículo presenta tres métodos que han demostrado ser útiles en los trabajos de investigación con estas poblaciones estigmatizadas, clandestinas o de difícil acceso: 1) Muestreo basado en sitios y horarios específicos (Venue Based Time- Location Sampling). 2) Muestreo de áreas en las que se localiza la población blanco o población objeto de estudio (Targeted Sampling). 3) Muestreo dirigido por el participante (Respondent Driven Sampling). Luego de analizar las características, ventajas y limitaciones de cada método de muestreo, se presenta una comparación de todos ellos en términos de validez externa, posibilidad de obtener muestras probabilísticas y empleo de investigación etnográfica.

Palabras clave: salud pública, métodos de muestreo, poblaciones ocultas.

[1] Dirección de correspondencia:
Email: Bernardo.Useche@uth.tmc.edu
[2] Email: marceq@javerianacali.edu.co

Abstract

Research projects that need to recruit "hard-to-reach" or "hidden populations" require sampling methods that not only facilitate data collection and confidentiality but also increase external validity and allow for statistically appropriate inferences. This paper presents three sampling methods useful for researchers who work with stigmatized hard-to-reach or clandestine populations: 1) Venue Based Time/ Location Sampling, 2) Targeted Sampling; and 3) Respondent Driven Sampling. The characteristics, advantages and limitations for each method are analyzed. They are also compared in terms of the possibility of obtaining probabilistic samples, external validity, and use of ethnographic research.

Key words: public health, sampling methods, hidden populations, venue based time- location sampling, targeted sampling, respondent driven sampling.

Resumo

Projetos de pesquisa que necessitam recrutar participantes pertencentes a povoações "oculta" ou "difíceis de encontrar" requerem de métodos de amostragem que não só facilitem a colheita dos dados e a confidencialidade mas também que aumentem a validade externa e permitam fazer inferências estatísticas apropriadas. Este artigo teórico apresenta três métodos que demonstraram ser úteis nos trabalhos de pesquisa com estas povoações estigmatizadas, clandestinas ou de difícil acesso: 1) Amostragem baseado em lugares e horários específicos (Venue Based Time- Location Sampling); 2) Amostragem de áreas nas quais se localiza a população branco ou população objeto de estudo (Targeted Sampling); e 3) Amostragem dirigido pelo participante (Respondent Driven Sampling). Depois de analisar as características, vantagens e limitações de cada método de amostragem, se apresenta uma comparação de todos eles em termos de validade externa, possibilidade de obter amostras probabilísticas e emprego de pesquisa etnográfica.

Palabras chave: Saúde Pública, os métodos de amostragem, Hidden Populações, Local Baseada Time-Local de amostragem, amostragem orientada, Reclamado Driven Amostragem.

Introduction

In the field of public health, the interest in developing sampling methods aimed at reaching "hidden populations" have been steered by the fact that monitoring risk-taking behaviors of stigmatized or clandestine subgroups such as injection drug users [IDUs], Men who have Sex with Men (MSM), and transgender sex workers is crucial for the epidemiologic surveillance of the HIV/AIDS epidemics (Lansky, Sullivan, Gallagher y Fleming, 2007a; Lansky, Abdul-Quader, Cribbin, Hall, Finlayson, Garfein, et al., 2007b).

While in the United States, The Center for Disease Control [CDC] has

systematically and successfully used Venue-Based Time-Location [VBTLS] sampling methods with that purpose since 1994 (MacKellar, Gallagher, Finlayson, Sanchez, Lansky, Sullivan, 2007; MacKellar, Valleroy, Karon, Lemp y Janssen, 1996), the introduction of Respondent Driven Sampling [RDS] ten years ago (Heckathorn, 1997; Heckathorn D. D., Broadhead R. S., Anthony D. L. y Weakliem D. L.,1999) was a significant contribution to the efforts for developing a more efficient recruitment of subjects in hard to reach populations as well as a methodology that sought to improve the external validity of the samples. A third method, Targeted Sampling [TS] which underscores the importance of extensive ethnographic work has also being used since the late 1980s as an "option for the study of hidden populations" (Watters y Biernacki, 1989). This paper describes and compares these three sampling methods (VBTLS, TS, y RDS) which to the best of our knowledge have not had received attention among public health and health psychology researchers in Latin America.

Other methods such as convenience and traditional snowball sampling are commonly used when studying hard to reach populations but they are not included in this analysis because they are better known sampling techniques and because of their limitations for producing unbiased estimates. In addition, one of the authors of this paper (BU) had the opportunity of working from 2003 to 2005 under the direction of Dr. Jan Risser in the implementation of Venue Based Time Location [4], Targeted Sampling, and Respondent Driven Sampling (Lansky et al., 2007; Robinson, Risser, McGoy, Becker, Rehman, Jefferson et al., 2006) methodologies for the CDC's National HIV Behavioral Surveillance [NHBS] project (Gallagher, Sullivan, Lansky y Onorato, 2007).

Venue based time – location sampling

Duncan Mackellar and his colleagues at the CDC (MacKellar et al., 2007) describe three basic components of venue based sampling: Formative research in order to produce the information needed to develop a universe of the venues located in the geographical area of the study; periodic elaboration of sampling frames for venues and Venues-Day-Time [VDTs] to guide data collection; and finally, conducting recruitment and interviewing in accordance to VDTs.

Figure 1, taken from Mackellar et al. (2007, p. 42) illustrates the type of sampling frames created when applying venue based sampling methodology. One of the main advantages of venue based sampling is that it allows rigorous planning of the sampling events which facilitates procurement of a probability sample of the visits and the generalization of the results to the studied population who attended the venues included in the sampling frames. However as Mackellar et al. (2007) have pointed out, it would be more important if the results can be generalized directly to the visitors which requires researchers to develop "and validate a weighting mechanism that uses venue attendance data measured in the survey to estimate a participant's selection probability" (p. 46). But even if such a weighting mechanism is constructed, some external validity problems will persist given that by definition venue based sampling does not reach the members of the target population who do not visit the venues included in the study.

Venues	VDTs						
Venue IDb	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
F001	6p-10p		6p-10p				
X002			8p-12a	8p-12a	8p-12a	8p-12a	
C019		6p-10p	6p-10p	6p-10p	6p-10p	6p-10p	4p-8p
					10p-12a	10p-12a	
P007						2p-6p	4p-6p
D101					11:30p 3:30a		
R045	6p-10p	6p-10p	6p-10p	6p-10p	6p-10p	6p-10p	
S033	4p-8p 8p-12a 12a-2a	4p-8p 8p-12a 12a-2a	4p-8p 8p-12a 12a-2a	4p-8p 8p-12a 12a-2a	4p-8p 8p-12a 12a-2a	4p-8p 8p-12a 12a-2a	4p-8p 8p-12a 12a-2a
D052			8p-12a	8p-12a	8p-12a	8p-12a	
O004			8p-9p				
O008		Tuesday 7p-10p (1st and 3rd)					
Z001	8p-12a						
X021	6p-10p 10p-2a	6p-10p 10p-2a	6p-10p 10p-2a	6p-10p 10p-2a	6p-10p 10p-2a	2p-6p 6p-10p 10p-2a	2p-6p 6p-10p 10p-2a
S001	6p-10p	6p-10p	6p-10p	6p-10p	6p-10p	6p-10p	4p-8p
					10p-2a	10p-2a	6p-10p
C001	6p-10p	6p-10p	6p-10p	6p-10p	8p-12a	8p-12a	

Figure 1. Hypothetical sampling frame of MSM-identified venues and associated venue-specific, day time periods (VDT's)a. Taken from Mackellar et al. (p. 42)

VDTs are venue-specific, day time periods expected to yield a minimum of eligible MSM.

[b] B= bar; C= café or restaurant; D= dance club; F= fitness club or gymnasium; G= Gay Pride or similar event; H= house party; O= social organization; P= park or beach (not public sex environment); R= retail business; S= street location (e.g. corner); V= rave, circuit park, or similar event; X= sex establishment or environment; Z= other.
MSM: men who have sex with men.

The same authors found that the application of this type of sampling method implies three methodological challenges: 1) Appropriate staff able to cope with the demanding circumstances of working unusual schedules, obtaining access to the venues, and dealing with the adverse environmental conditions while collecting

high quality data. 2) Community support from organizations representative of the studied population and venue owners and managers; and 3) Ongoing formative research in order to keep identifying new venues or venues programming seasonal events.

Targeted Sampling

Targeted sampling requires extensive formative research and complex ethnographic work. In order to identify the locations where the study participants will be recruited, is recommendable that professional outreach workers and experienced ethnographers map the areas based on field observations and interviews with key informants from inside and outside the subgroups which will be studied. Exhaustive review of the secondary data is important for characterizing the target population.

Based on information from the formative research, ethnographic mapping and a sampling frame of locations are developed. In the first IDUs cycle of the NHBS in Houston, the formative research conducted produced a reasonable list of outdoor places and street corners (Robinson, et al., 2006; Risser, Useche y Rehman, 2005). Finally, staff members select randomly from the sampling frame the locations where participants will be recruited and interviewed.

The map in Figure 2, taken form the report on the formative research conducted for the first NHBS IDU cycle in Houston, (Risser et al., 2005. p.6) shows the distribution of noticeable drug use in Houston (represented by the places where drug arrests occurred) and the areas of special interest for recruitment identified during field observations. When compared the distribution of drug use with the distribution of HIV cases with IDU as a risk factor, and with the distribution of Hepatitis C cases, the resultant maps were pretty much the same. Even more interesting from the perspective of social epidemiology, these maps overlap with the poverty map of Houston.

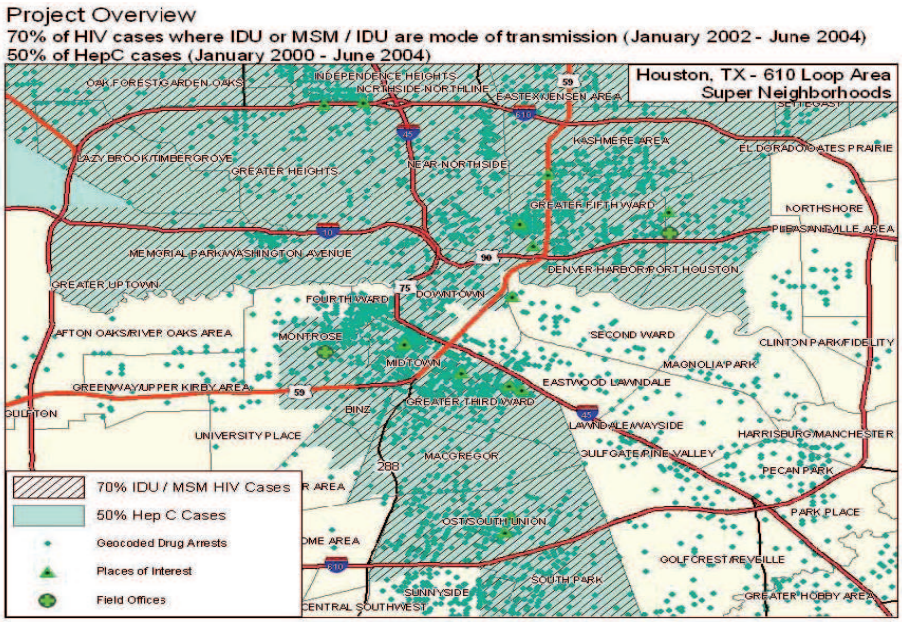

Figure 2. Taken from Risser, Useche, Rehman et al., 2005. (p. 6)

The ethnographic data collected is one of the strengths of targeted sampling. It provides a previous characterization of the population, a better knowledge of the socioeconomic determinants of risk-taking behaviors, and key information about personal situations and motivational factors associated with the variables studied. In this context, recruitment is not only matter of selecting members to participate in the study but it also involves a deeper exploration of the individuals, their networks and their communities.

Although this method does not produce unbiased samples and its external validity is very low because generalization of the results is limited to the population actually reached through the recruitment process, targeting sampling could be the choice of preference for research aimed to design and develop prevention interventions.

Besides the problems with external validity, Semaan, Lauby and Libman (2002) enumerate the following other two limitations of targeted sampling: frequently secondary data does not provide all the information expected. For example, data on population characteristics reported by zip codes does not include information on population differences within a same zip code area. Second, because of the characteristics of the recruitment, the sample obtained is difficult to replicate and interviewer biases are common. For example, security concerns may limit the recruitment to certain areas considered to be safer.

Additionally, as it was observed during the pilot study for the first IDU cycle of NHBS (Robinson et al., 2006), the presence of project staff causes curiosity as well as defensive reactions among members of the community in the data collection area; poor weather conditions affects recruitment negatively, and the fact that recruitment and interviews are conducted in the field makes it difficult to keep privacy and avoid interferences. Properly addressing these situations usually implies extra costs (such as renting an interview van and hiring more staff).

The same study found that locations initially identified as high yield recruitment areas do not provided the expected number of eligible participants which requires spending more time to recruit the planned number of subjects or the mobilization to other previously selected areas of recruitment.

Respondent driven sampling

Respondent Driven Sampling (Lansky et al., 2007a; Heckathorn, 1997; Heckathorn y Jeffri, 2003; Heckathorn, Semaan, Broadhead y Hughes, 2002) is a relatively new method to sample hard-to-reach populations and a sophisticated variant of snowball sampling. Both methods rely on a chain referral strategy where the initial participants or "seeds" refer their peers who in turn should do the same and so on. But in contrast with snowball sampling which produces nonprobability samples biased because of the selection of "seeds" (Erickson, 1979), RDS can produce unbiased samples that are independent of the initial recruiters (Lansky et al., 2007b; Heckathorn, 2002; Magnani, Sabin, Saidel y Heckathorn, 2005) while keeping the capacity of networks for reaching the most diverse members of the hidden population and hence providing a more complete coverage of the target population.

In RDS a few "seeds" are chosen by the researchers based on their potential for recruiting eligible subjects. In order to control the seeds' and their referrals' trend to recruit differentially people like themselves among their network contacts, RDS uses a coupons system to limit the number of

recruits per participant; and based on Watts' "small worlds" theory (Watts, 2003) allows continuing the recruitment for several "waves".

An important feature of RDS is the introduction of specific software: a Coupon Manager [RDSCM] and an Analysis Tool [RDSAT]. The coupon manager permits researchers to track the actual recruitment carried out by each individual, the link recruiter-recruit, and the network size for each participant. With this information and using RDSAT populations estimates and recruitment matrices can be calculated (Lansky, 2007a).

The main strength of RDS resides on its theoretical, methodological, and statistical basis, an issue beyond the purposes of this paper. But it is clear that the ongoing work by Heckathorn and his colleagues (Heckathorn, 2007) is contributing to the advancement of sampling methodology for hard to reach or hidden populations.

Figure 3 taken from Lansky et al. (2007a, p. 51) describes the sequence of the RDS recruitment process which starts with the critical step of selecting the "seeds" or initial recruiters and it ends with giving the incentives to the recruiters for the referrals of eligible participants who have been already interviewed.

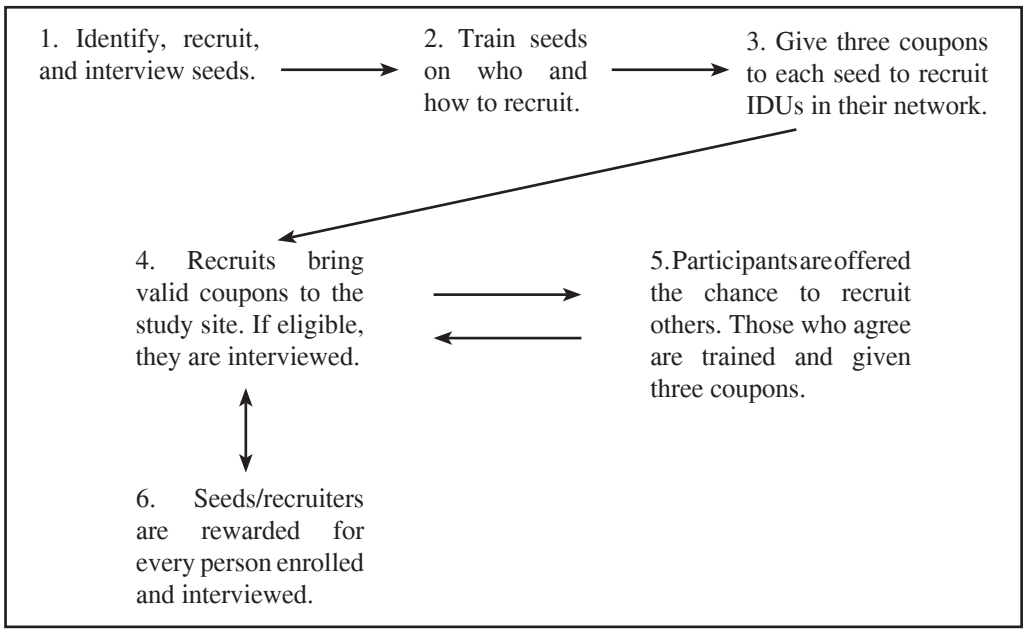

Figure 3. RDS Recruitment Methods. Taken from Lansky et al. (2007a, p. 51)

A pilot study comparing the application of RDS and TS for the first NHBS IDU cycle has been conducted (Lansky et al., 2007a; Robinson et al., 2006). Its conclusions as well as other observations have been incorporated to the comparative table on Venue Based, Targeted Sampling and Respondent Driven Sampling shown in Table 1.

Table 1. Comparing Targeted Sampling, Venue Based and Respondent Driven Sampling

Sampling Method	Probability Sample	External Validity	Ethnographic / Formative Research	Recruiters	Working Environment	Cost / Efficiency Preliminary conclusions More research needed
Targeted Sampling	No. It does not allow calculate population estimates	Limited	Extensive and complex. It provides valuable information on socioeconomic determinants of risk behaviors	Professional outreach workers familiar with the target population are recommended.	TS is conducted in the field: Outdoors / Street.Safety concerns. Weather conditions affect recruitment. Lack of privacy	It performs satisfactorily but it requires more staff-time per recruit. High costs because of larger person-hours expenditures.
Venue based / time-location sampling	Yes. "it produces a probability sample of visits to venues included within sampling frames"	Considerable. There is always risk of sampling bias.	Extensive. Formative research is required in order to identify venues, attendance patterns, times, and recruitment methods. Ongoing formative research is required	Usually the same staff in charge of data collection. Recruiters must be highly motivated and receive specific training for conducting sampling events.	Usually, recruitment and data collection is conducted late at night. Participants can be interviewed inside venues; in an interview van; or in a place nearby. Smoking and other poor / risky environment conditions existent in many venues generate health and safety concerns among data collectors.	It performs satisfactorily. Costs information is not available.
Respondent driven sampling	Yes It allows to calculate population estimates. RDS produces "asymptotically unbiased population estimates when its assumptions are satisfied". The sampling method is unbiased for samples of meaningful size.	Considerable. RDS "extends the sample to all potential members of a subgroup selected for surveillance by accessing respondents through their social networks"	Minimal. Aimed to identify the "seeds". The data collection itself provides information that is useful as ongoing formative research.	"Seeds" and peers referral through their own network. Selection of seeds is critical. "research hustler" seeds are not good recruiting eligible participants.	Data collection requires of a storefront with separated areas for waiting, coupon management, and conducting interviews. It offers more control over the data collection process y better safety conditions for the staff.	It performs satisfactorily. In some sites implies higher costs than TS because of store front rent; need of a computer, and incentive payouts for referrals.

Conclusion

This article has described the characteristics, advantages and limitations of Targeted, Venue Based, and Respondent Driven Sampling methods in the context of their current use in the United States when conducting research with "hidden populations". Although the three methods

have performed satisfactorily, a comparison of the three shows that researchers should carefully consider the objectives and resources of their project when choosing a sampling method. Targeted Sampling requires intensive ethnographic work in order to locate the places to recruit participants and plan data collection. That ethnographic data will also provide important qualitative information on the socioeconomic context and living conditions of the studied population. However, Targeted Sampling does not permit the calculation of population estimates. Venue based sampling allows a very systematic recruitment of participants; requires the collaboration of the venue managers, and produces probability samples but only of the visits to the venues included in the sample. Respondent Driven Sampling is based on participants' social networks, requires careful selection of initial "seeds" and permits calculation of population estimates.

References

Erickson B. H. (1979). Some problems of inference from chain data. *Sociological Methodology*, *10*, 276-302.

Gallagher K. M, Sullivan P. S, Lansky A. y Onorato I. M. (2007). Behavioral surveillance among people at risk for HIV infection in the U.S.: the National HIV Behavioral Surveillance System. *Public Health Report, 122*, 32-8.

Heckathorn D. D. y Jeffri J. (2003). "Social Networks of Jazz Musicians." Pp. 48-61 *in Changing the Beat: a study of the worklife of jazz musicians.* Volume III: Respondent-Driven Sampling: Survey Results by the Research Center for Arts and Culture, National Endowment for the Arts Research Division Report #43. Washington, DC.

Heckathorn D. D. (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems, 44*, 174-99.

Heckathorn D. D., Broadhead R. S., Anthony D. L. y Weakliem D. L. (1999). AIDS and social networks: prevention through network mobilization. *Sociological Focus, 32*,159-79.

Heckathorn, D. (2007). Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment. *Sociological Methodology*. In Press. Available online: http://www.respondentdrivensampling.org/

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

Heckathorn, D. D., Salaam S., Broadhead R. S. y Hughes J. J. (2002). Extensions of respondent-driven sampling: a new approach to the study of injection drug users aged 18–25. *AIDS and Behavior*, *13,* (1), 55-67.

Lansky A, Abdul-Quader A. S., Cribbin M., Hall T., Finlayson T. J., Garfein R. S., Lin, L. S. y Sullivan, P. S. (2007). Developing an HIV behavioral surveillance system for injecting drug users: the National HIV Behavioral Surveillance System. *Public Health Report, 122*, 48-55.

Lansky, A., Sullivan, P. S., Gallagher, K. M. y Fleming, P. (2007). HIV Behavioral Surveillance in the U.S.: A Conceptual Framework. *Public Health Report, 122*, 16-23.

MacKellar D. A., Gallagher K. M., Finlayson T., Sanchez T., Lansky A. y Sullivan P. S. (2007). Surveillance of HIV risk and prevention behaviors of men who have sex with men-a national application of venuebased, time-space sampling. *Public Health Report, 122*, 39-47.

MacKellar D., Valleroy L., Karon J., Lemp G. y Janssen R. (1996). The Young Men's Survey: methods for estimating HIV seroprevalence and risk factors among

young men who have sex with men. *Public Health Report, 111,* 1, 138-44.

Magnani R., Sabin K., Saidel T. y Heckathorn D. D. (2005). Review of sampling hard-to-reach and hidden populations for HIV surveillance. *AIDS, 19* (2), S67-72.

Risser, J., Useche, B., Rehman, H. (2005). *National HIV Behavioral Surveillance among Injecting Drug Users: Final Formative Assessment Report.* Unpublished manuscript, The University of Texas Health Science Center at Houston, School of Public Health and Department of Health and Human Services Bureau of Epidemiology Houston, Texas.

Robinson W. T, Risser J., McGoy S., Becker A., Rehman H., Jefferson M., Griffin V., Wolverton M. y Tortu S. (2006). Recruiting injection drug users: A three-site comparison of results and experiences with respondent-driven and targeted sampling procedures. *Journal of Urban Health, 83* (7), 129-138.

Semaan, S., Lauby, J. y Liebman, J. (2002). Street and Network Sampling in Evaluation Studies of HIV Risk-Reduction Interventions. *AIDS Review, 4*,213-223.

Watters J. K. y Biernacki P. (1989). Targeted sampling: options for the study of hidden populations. *Social Problems,* 36 (4), 416-430.

Watts, D. J. (2003). Small Worlds: *The Dynamics of Networks between Order and Randomness*. Princeton, NJ: Princeton University Press.