Sampling bias due to structural heterogeneity and limited internal diffusion

Jukka-Pekka Onnela[1,2,3,*] Neil F. Johnson[4], Sean Gourley[1,2], Gesine Reinert[5], and Michael Spagat[6]
[1] *Department of Physics, University of Oxford, Oxford OX1 3PU, UK*
[2] *CABDyN Research Cluster, Said Business School, University of Oxford, Oxford, OX1 1HP, UK*
[3] *DBEC, Helsinki University of Technology, P.O. Box 9203, FIN-02015 HUT, Finland*
[4] *Physics Department, University of Miami, Coral Gables, Florida 33124, USA*
[5] *Department of Statistics, University of Oxford, Oxford OX1 3TG, UK*
[6] *Department of Economics, Royal Holloway, University of London, TW20 0EX, UK*
(Dated: July 25, 2008)

Complex systems research is becomingly increasingly data-driven, particularly in the social and biological domains. Many of the systems from which sample data are collected feature structural heterogeneity at the mesoscopic scale (i.e. communities) and limited inter-community diffusion. Here we show that the interplay between these two features can yield a significant bias in the global characteristics inferred from the data. We present a general framework to quantify this bias, and derive an explicit corrective factor for a wide class of systems. Applying our analysis to a recent high-profile survey of conflict mortality in Iraq suggests a significant overestimate of casualties.

PACS numbers: 89.65.-s, 89.75.Fb, 89.75.-k

Monitoring large social or biological systems bears similar challenges to monitoring many-particle systems in physics. The increasing availability of data on human behaviour from information and communication technologies [1, 2] and data from high throughput techniques in biology enable scientists to study these diverse systems with similar methodologies. Many biological or social systems are not internally homogeneous, but instead feature time-dependent community groupings and limited inter-community mixing [1–4]. Individuals form dynamic groups in professional and private settings reflected in, for example, structures of scientific collaboration and mobile phone call patterns [3]. The cell nucleus consists of multiple compartments with different micro-environments that exist in spatially localised regions in the heterogeneous intranuclear space [4]. In this Letter, we quantify the consequences of sampling a subset of objects in such a system. Starting with a general theoretical framework, we show that the interplay between heterogeneity and limited diffusion can yield a substantial bias in the inferred global characteristics. We consider the specific example of a recent conflict mortality study in Iraq, and find support for a significant positive bias in the inferred casualty numbers.

Consider a large system made up of N particles characterised by a microscopic state variable x_i. The system is heterogeneous in that it consists of m different subsystems or communities $S_1, \ldots, S_m$ with N_i particles in S_i such that $N_1 + \ldots + N_m = N$. The subsystems are interconnected in some limited way, thereby allowing for only partial diffusion or mixing of particles between them. We wish to learn about the state of the system described by the extensive macroscopic variable $X = \sum_{i=1}^{N} x_i$ but, in line with typical empirical scenarios, assume that we cannot observe the entire system. Instead, we monitor the state of a set of tagged particles in different subsystems and use this data to make statistical inferences about X.

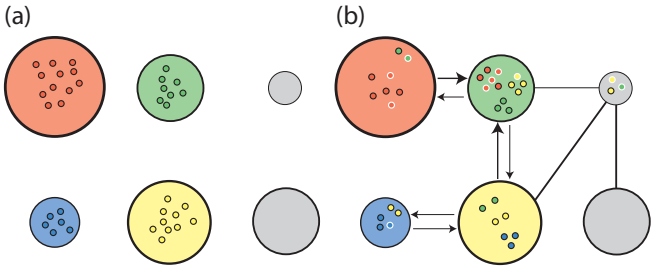

FIG. 1: (Color online) (a) The system is prepared by tagging some particles in some of the subsystems, which corresponds to a sampling process. The particles are non-interacting and indistinguishable apart from the initial subsystem given by their color. (b) After the initial state, the matrix $\mathbf{f} \neq \mathbf{I}$ quantifies mixing between subsystems. It can be interpreted as a weighted and directed network adjacency matrix of the subsystems. The state of particle $x_i \in \{0, 1\}$ is indicated by coloring its circumference black or white, respectively. Only tagged particles are visible and available for analysis.

Let us assume that the particles are identical and non-interacting and that each can be in one of two states $x_i \in \{0, 1\}$. The system is initially prepared with $x_i = 0$ for all i and only irreversible $0 \to 1$ changes are considered. Microscopic state changes are subsystem specific, with the element q_k of the vector $\vec{q}$ specifying the probability for a particle in subsystem S_k to change state. Hence the x_i are independent random variables and within a given subsystem are identically distributed; denote by y_k a random variable having the distribution of any x_i present in S_k. The state of a particle can be identified with, for example, the staining of cancerous cells in a biological organism under medical imaging (stained vs. clear), or the disease status of an individual (healthy vs. diseased). The subsystem specific probabilities $\vec{q}$ could arise from there being different numbers of cancerous cells or pathogens in these systems.

The mixing of particles is governed by the constant mixing matrix $\mathbf{f} = [\vec{f}_1 \vec{f}_2 \cdots \vec{f}_m]$, where $\vec{f}_i$ specifies the fraction of time particles initially placed in S_i spend in other subsystems. The entries of $\mathbf{f}$ can be interpreted as probabilities of finding particles in different subsystems (see Fig. 1). The diagonal elements f_{ii} correspond to the probability of finding a particle in its initial subsystem. Note that $\mathbf{f}$ does not need to be symmetric. In the limit as the mobility of the particles tends to zero, the matrix $\mathbf{f}$ consists of only diagonal elements $f_{ij} = \delta_{ij}$, with the effect that the subsystems become completely isolated.

Denote by X_i is the contribution of all particles initially in S_i towards X, and by X_{ij} the contribution of a single particle $j = 1, \ldots, N_i$ initially in S_i towards X_i. Let $D_{i,k}$ denote the number of particles initially in S_i which are observed in S_k; then $D_{i,k}$ follows a binomial distribution with parameters N_i and f_{ik}. We write our quantity of interest X as

$$X = \sum_{i=1}^{m} X_i = \sum_{i=1}^{m} \sum_{j=1}^{N_i} X_{ij} = \sum_{i=1}^{m} \sum_{k=1}^{m} \sum_{\ell=1}^{D_{i,k}} y_{k,\ell}, \quad (1)$$

where the $y_{k,\ell}$ are independently distributed as y_k. Consider now a situation where, to estimate X, we can draw samples from only some of the subsystems. Let S consisting of m subsets S_k denote the set of all subsystems and let $S' = \bigcup_{k=1}^{m'} S_k$, i.e. the first m' of these sets, denote the set of samplable subsystems. The expectation value of X in the entire system, $\langle X \rangle_S$, and in the samplable system, $\langle X \rangle_{S'}$, is given by

$$\begin{aligned} \langle X \rangle_S &= \sum_{i=1}^{m} \langle X_i \rangle = \sum_{i=1}^{m} \sum_{k=1}^{m} \langle D_{i,k} \rangle \langle y_k \rangle = \sum_{i=1}^{m} N_i \sum_{k=1}^{m} f_{ik} \langle y_k \rangle \\ \langle X \rangle_{S'} &= \sum_{i=1}^{m'} \langle X_i \rangle = \sum_{i=1}^{m'} N_i \sum_{k=1}^{m} f_{ik} \langle y_k \rangle, \end{aligned} \quad (2)$$

respectively. If the subsystems are heterogeneous and this is not accounted for in the sampling procedure, we may incur a significant bias. To quantify this, we define the *bias factor* R as the scaled ratio of $\langle X \rangle_{S'}$ to $\langle X \rangle_S$,

$$R = \frac{\sum_{i=1}^{m'} N_i \sum_{k=1}^{m} f_{ik} \langle y_k \rangle / N'}{\sum_{i=1}^{m} N_i \sum_{k=1}^{m} f_{ik} \langle y_k \rangle / N} = \frac{N \sum_{i=1}^{m'} \eta_i}{N' \sum_{i=1}^{m} \eta_i}, \quad (3)$$

where $N' \leq N$ is the number of particles in S' and values of $R > 1$ ($R < 1$) correspond to overestimating (underestimating) the expectation value of X in the system when sampling is based on subsystems in S' only. The contribution of a set of identical particles initially located in subsystem j (from subsystem k), denoted by η_j (η_{jk}), towards the above expectation value is given by

$$\eta_i = \sum_{k=1}^{m} \eta_{ik} = N_i \sum_{k=1}^{m} f_{ik} \langle y_k \rangle = \langle X_i \rangle. \quad (4)$$

A special case of the framework arises when the microscopic state variables x_i and y_k correspond to independent Bernoulli trials related to some event ω. We assume that the event ω occurs independently of the mixing. Now q_i $(1 - q_i)$ is the probability of observing $x = 1$ ($x = 0$) in subsystem S_i long enough after the initial state so that the system has reached an equilibrium. Regardless of the number of subsystems present, the system can always be divided into a samplable subsystem and a non-samplable subsystem. Let $S_I = S'$ and let the remaining subsystems form the non-samplable subsystem $S_O = \bigcup_{k=m'+1}^{m} S_k$. As a mnemonic, the subscript I refers to in-sample and O to out-of-sample. Note that whereas before $S' \subseteq S$, here $S_I \cap S_O = \emptyset$. We now have $N_I = N' = \sum_{k=1}^{m'} N_k$ and $N_O = \sum_{k=m'+1}^{m} N_k$, corresponding to the number of particles in S_I and S_O, respectively, and $N_I + N_O = N$. We define the 'renormalised' probabilities $q_I = N_I^{-1} \sum_{k=1}^{m'} N_k q_k$ for a particle to be subjected to ω while present in S_I and its complement $1 - q_I$ for the particle to not be subjected to ω while present in S_I. Similarly, we define for S_O the probability $q_O = N_O^{-1} \sum_{k=m'+1}^{m} N_k q_k$ (and its complement $1 - q_O$) for a particle to (not) be subjected to ω while present within S_O. Finally, we define the mobility factors such that f_I (f_O) is the probability for a particle initially placed in S_I (S_O) to be present within S_I (S_O), and $1 - f_I$ ($1 - f_O$) is the probability for a particle initially placed in S_I (S_O) to not be present within S_I (S_O), i.e., to be present within S_O (S_I). These are written as

$$\begin{aligned} f_I &= N_I^{-1} \sum_{i=1}^{m'} N_i \sum_{j=1}^{m'} f_{ij} \\ f_O &= N_O^{-1} \sum_{i=m'+1}^{m} N_i \sum_{j=m'+1}^{m} f_{ij}. \end{aligned} \quad (5)$$

We now define $\pi_{\alpha\beta}$ with $\alpha, \beta \in \{O, I\}$ as the probability that a particle picked uniformly at random was placed initially in S_α with $x_i = 0$ and changes state to $x_i = 1$ in S_β. This leads to $\pi_{OO} = \frac{N_O}{N_I + N_O} f_O \, q_O$, $\pi_{OI} = \frac{N_O}{N_I + N_O} (1 - f_O) \, q_I$, $\pi_{IO} = \frac{N_I}{N_I + N_O} (1 - f_I) \, q_O$, and $\pi_{II} = \frac{N_I}{N_I + N_O} f_I \, q_I$. The sum $\pi_{II} + \pi_{IO} + \pi_{OI} + \pi_{OO}$ is the probability that a randomly chosen particle is subjected to ω and hence changes its microscopic state. The expected number of particles with $x_i = 1$ in a population of size N is hence $N_O \, f_O \, q_O + N_O \, (1 - f_O) \, q_I + N_I \, (1 - f_I) \, q_O + N_I \, f_I \, q_I = (q_I - q_O)(f_I N_I - f_O N_O) + q_I N_O + q_O N_I$, whereas the probability that a randomly chosen particle in S_I changes state is $q_I f_I + q_O (1 - f_I)$. Hence the expected number of realizations for a population of size N, based on the rate for S_I only, would be $(N_I + N_O)[q_I f_I + q_O (1 - f_I)]$. We obtain

$$R = \frac{(N_I + N_O)[q_I f_I + q_O (1 - f_I)]}{(q_I - q_O)(f_I N_I - f_O N_O) + q_I N_O + q_O N_I}. \quad (6)$$

Assuming that $N_I \neq 0$ and $q_O \neq 0$, and setting $q = q_I/q_O$ and $n = N_O/N_I$, we obtain

$$R = R(f_I, f_O, q, n) = \frac{(1+n)(1+qf_I - f_I)}{(q-1)(f_I - f_O n) + qn + 1}. \quad (7)$$

Hence the bias factor R depends only on f_I, f_O, and the ratios $q = q_I/q_O$ and $n = N_O/N_I$. Finally, in the case of symmetric mobility $f_I = f_O = f$ the above simplifies to

$$R = R(f, q, n) = \frac{(1+n)(1+qf - f)}{f(q-1)(1-n) + qn + 1}. \quad (8)$$

The no-bias limit of $R = 1$ requires either (1) $n = 0$ (i.e. $N_O = 0$) implying that no particle is placed initially in S_O, or (2) $q = 1$ (i.e. $q_I = q_O$) implying equal rates of changing state in S_I and S_O, or (3) $f = 1/2$ which suggests that particles based in S_I spend on average half of their time in S_O and vice versa. Setting $R(f, q, n) = r$ for general r and solving for q in terms of n and f yields

$$q(f, n, r) = \frac{f(1+n+nr-r) + r - n - 1}{f(1+n+nr-r) - nr}. \quad (9)$$

Although q is unobservable, we can estimate $\tilde{q} = N^{-1}\sum_{i,j} X_{ij}$ and $\tilde{q}' = (N')^{-1}\sum_{i=1}^{m'}\sum_{j=1}^{N_i} X_{ij}$, leading to the asymptotically unbiased estimator $\hat{R} = \tilde{q}'/\tilde{q}$ for the bias factor R. If $R = 1$ then we would expect that $\hat{R} \approx 1$. The variation in $\hat{R}$ can be assessed via a normal approximation [7]. Basing $\tilde{q}'$ on S_I and assuming that $\langle X \rangle_S$ is not too small, the approximate variance is

$$\mathrm{Var}(\hat{R}) \approx \frac{(1+n)^2\{fq(1-q_I) + (1-f)(1-q_O) + f(1-f)(qq_I + q_O - q)\}}{q_0 N_I[f(q-1)(1-n) + qn + 1]^2}. \quad (10)$$

We will now exemplify the above framework by applying it to study conflict mortality. To estimate the number of casualties in a conflict, one would ideally like to have access to a complete national list of households from which a sample could be drawn at random. Even when this scenario is feasible, the selected households are widely scattered, which is costly not only in terms of time and money, but also exposes the researchers to high levels of risk. To overcome these concerns, recent studies economise resources by using a cluster sampling methodology. This hierarchical sampling process involves making choices on how to choose large geographic areas and how to proceed from them to individual households.

We can equate particles in the framework with individuals such that the system size N corresponds to the population of the country and the state of each particle $x_i \in \{0, 1\}$ corresponds to the individual being alive or dead (where the death has resulted from conflict related violence), respectively. The different subsystems correspond to heterogeneous areas that are characterised by varying levels of violence such that the probability for an individual to be killed in S_k is given by q_k. Note that these areas, or zones, may be fragmented and interdispersed. Now $\langle X_k \rangle$ corresponds to the expected number of casualties in S_k for a given q_k, and $\langle X \rangle$ corresponds to the expected number of casualties in the country. Daily human movement between different areas is quantified by the mixing matrix. The initial subsystem of a particle can be identified with the residential zone of the individual. The 'renormalised' systems S_I and S_O correspond to sets of subsystems that may or may not be sampled, respectively, given the sampling method. To include an individual in the study, his or her home needs to be located in the samplable subsystem S_I.

FIG. 2: Part of an original hand drawn map showing the clusters of cholera cases in the London epidemic of 1854. The crosses indicate wells and the dots infected households. Adapted from Ref. [8].

Here we focus on the final stages of the sampling procedure that was used estimate conflict mortality in Iraq [5], and we refer to it as the *Cross Street Sampling Algorithm (CSSA)*: (1) Select a "constituent administrative unit", (2) select a main street from "a list of all main streets",

(3) select randomly a residential street from "a list of residential streets crossing the main streets", (4) enumerate the households on the street, (5) select one household at random to initiate the interviewing, proceeding to 39 further adjacent households.

Compare this scenario to the famous case of identifying a public water pump on Broad Street (Fig. 2) as the cause of a cholera outbreak [8]. The individuals residing in the area served by the faulty pump have a high risk of contracting cholera. Therefore, within this "subsystem", the location of people and the location of pathogens are strongly correlated. For conflicts like the one in Iraq, violent events tend to be focused around cross-streets since they are a natural habitat for patrols, convoys, police stations, parked cars, roadblocks, cafes and street-markets [6]. Because of the prevalence of violence, mixing of populations between the zones is minimal. Because the cross-streets that are chosen for sampling, the location of violence and the location sampled sites are correlated by means of accessibility. Hence the micro-level details on household selection are crucial.

The parameter $n = N_O/N_I$ gives the proportion of population resident in S_O to that resident in S_I. Street layouts in Iraq are mostly irregular, hence CSSA will miss any neighbourhood not in the immediate proximity of a cross-street. Analysis of Iraqi maps suggests $n = 10$ is plausible [6]. The parameter $q = q_I/q_O$ gives the relative probability of death for anyone present in S_I, regardless of their zone of residence, to that of S_O. Given the extent and frequency of attacks, $q = 5$ is plausible [6]. The parameter $f = f_I = f_O$ gives the fraction of time spent by residents of S_I (S_O) in S_I (S_O). Given the nature of the violence, travel is limited; women, children and the elderly tend to stay close to home. Using the time people spend in their homes as a lower bound on the time they spend in their zones, assuming that there are two working-age males per average household of seven [5], with each spending 6h per 24h day outside their own zone, yields $f = f_I = f_O = 5/7 + 2/7 \cdot 18/24 = 13/14$ [6].

These values yield $R = 3.0$, suggesting that the Iraq estimate [5] provides a substantial overestimate of deaths by a factor of 3. To gauge the sensitivity of these results, we perform a simple sensitivity analysis by evaluating R for different values of parameters (Fig. 3). This shows the effect of relaxing the constraint $f = f_I = f_O$ and it is clear that in the limit of no mobility ($f_I = f_O = 1$) the bias is greatest. Conceptually speaking, the bias emerges from having simultaneously partial localization of violence and partial localization of people. Both of these conditions are needed for the bias to emerge, since if $f = 1/2$, we have $R = 1$ regardless of q and n, and if $q = 1$ we have $R = 1$ regardless of n and f.

We have presented a framework that can be used to gauge sampling bias in systems consisting of heterogeneous subsystems with limited inter-diffusion. The conclusions of our theory for a recent conflict mortality study [5] are supported by recent independent research [9–11]. Given the difficulty in providing representative samples in many social and biological systems, our theory should prove invaluable in correcting for resulting biases.

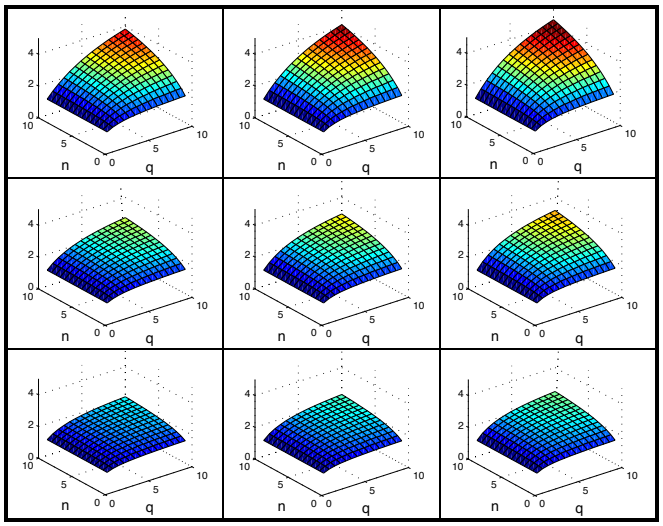

FIG. 3: (Color online) Sensitivity analysis of bias factor R defined by Eq. 7. Each panel shows $R = R(f_I, f_O, q, n)$ with the values of f_I and f_O fixed for each panel. Here f_I (f_O) varies by columns (rows) over the values $\{0.75, 0.85, 0.95\}$ increasing from left to right (bottom to top).

Acknowledgements: JPO acknowledges Wolfson College, Oxford. GR acknowledges MMCOMNET, Grant No. FP6-2003-BEST-Path-012999.

* Electronic address: jp.onnela@physics.ox.ac.uk

[1] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, *Proc. Natl. Acad. Sci USA* **104**, 7332 (2007).
[2] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabási, *Nature* **453**, 779 (2008).
[3] G. Palla, A.-L. Barabási, and T. Vicsek, *Nature* **446**, 664 (2007).
[4] R. Grima and S. Schnell (2008), *Essays in Biochemistry* **44**, in press (2008).
[5] G. Burnham, R. Lafta, S. Doocy, and L. Roberts, *Lancet* **368**, 1421 (2006).
[6] N. F. Johnson, M. Spagat, S. Gourley, J.-P. Onnela, and G. Reinert, *Journal of Peace Research* **45**, 653 (2008).
[7] V. Barnett, *Sample Surveys. Principles and Methods*, 2nd ed., Ewdard Arnold, London, 1991.
[8] J. Snow, *On the Mode of Communication of Cholera*, John Churchill, London, 1855.
[9] UN Development Program, *Iraq Living Conditions Survey 2004*, 2005.
[10] Iraq Body Count Project, *A Dossier of Civilian Casualties in Iraq 2003-2005*, 2005.
[11] Iraq Family Health Survey Study Group, *New England Journal of Medicine* **358**, 484 (2008).