CIMPA

Epi-SCIENCE
The Science Behind Population Health
https://revistas.ucr.ac.cr/index.php/episcience

Revista Interdisciplinaria Epi-SCIENCE

On the scaling patterns of infectious disease incidence in cities

Patterson-Lomba, Oscar[1], Andrés Gomez-Lievano[1,2]

[1] *Analysys Group, Inc. Boston, USA.*
[2] *Center for International Development, Harvard Kennedy School, Harvard University, Cambridge, MA, USA.*

oscar.patterson-lomba@analysisgroup.com

Submitted: 01/07/2022; Accepted: 01/02/2023
DOI: https://doi.org/10.15517/es.2023.55397

Abstract

Urban areas with larger and more connected populations offer an auspicious environment for contagion processes such as the spread of pathogens. Empirical evidence has revealed a systematic increase in the rates of certain sexually transmitted diseases (STDs) in urban areas with larger population sizes. With rampant urbanization worldwide, it is increasingly important to improve our understanding of the key drivers of these systemic infection patterns. Using confirmed-case data for three STDs in U.S. metropolitan areas, we investigate the scaling patterns of the incidence of these infectious diseases in urban areas. The most salient features of these patterns are that, on average, the incidence of infectious diseases that transmit with less ease i) scale more steeply with population size, ii) are more variable across cities of similar size, and iii) are less predictable across time. These features are explained using a simple mathematical model of contagion and also through the lens of a new theory of urban scaling. These theoretical frameworks help us reveal the links between the factors that determine the transmissibility of infectious diseases and the properties of their scaling patterns across cities.

Keywords: scaling patterns; incidence; infectious diseases; urban areas; models.

CIMPA

Epi-SCIENCE
The Science Behind Population Health
https://revistas.ucr.ac.cr/index.php/episcience

Revista Interdisciplinaria Epi-SCIENCE

Sobre los patrones de escala de la incidencia de enfermedades infecciosas en ciudades

Patterson-Lomba, Oscar[1], Andrés Gomez-Lievano[1,2]

[1] *Analysys Group, Inc. Boston, USA.*
[2] *Center for International Development, Harvard Kennedy School, Harvard University, Cambridge, MA, USA.*

oscar.patterson-lomba@analysisgroup.com

Enviado: 07/01/2022; Aceptado: 02/01/2023
DOI: https://doi.org/10.15517/es.2023.55397

Resumen

Las áreas urbanas con poblaciones más grandes y conectadas ofrecen un entorno propicio para procesos de contagio como la propagación de patógenos. La evidencia empírica ha revelado un aumento sistemático en las tasas de ciertas enfermedades de transmisión sexual (ETS) en áreas urbanas con mayor tamaño de población. Con la urbanización desenfrenada en todo el mundo, es cada vez más importante mejorar nuestra comprensión de los impulsores clave de estos patrones de infección sistémica. Usando datos de casos confirmados para tres ETS en áreas metropolitanas de EE. UU., investigamos los patrones de escala de la incidencia de estas enfermedades infecciosas en áreas urbanas. Las características más destacadas de estos patrones son que, en promedio, la incidencia de enfermedades infecciosas que se transmiten con menos facilidad i) aumentan más abruptamente con el tamaño de la población, ii) son más variables entre ciudades de tamaño similar, y iii) son menos predecibles entre ciudades. tiempo. Estas características se explican utilizando un modelo matemático simple de contagio y también a través de la lente de una nueva teoría de escalamiento urbano. Estos marcos teóricos nos ayudan a revelar los vínculos entre los factores que determinan la transmisibilidad de las enfermedades infecciosas y las propiedades de sus patrones de escala en las ciudades.

Palabras clave: patrones de escala; incidencia; enfermedades infecciosas; áreas urbanas; modelos.

1. Introduction

Urban populations throughout the world are growing rapidly, with predictions indicating that by mid-century, over 60% of the population will live in urban areas [59]. This explosion of urban dwellers makes it paramount to decipher how this urbanization transition can be sustainable and beneficial on a global scale. Crucial to this enterprise is to improve our understanding of how infectious diseases spread and evolve in cities, as well as the overall public-health implications of rampant urbanization [2, 3, 61, 1, 18, 12, 24, 34, 19, 4, 37].

By their nature, cities foster large populations living in high proximity, leading to high levels of human interaction and mobility. Cities are thus suitable landscapes for pathogens to spread locally and internationally [11, 41, 1, 6]. Rapid growth and densification of urban populations in developing countries, often combined with poor living conditions and a precarious public health infrastructure, offer a particularly favorable landscape for the spread of certain infectious diseases. For example, increased population density and urbanization may have played a critical role in the worldwide dissemination of HIV [49, 12, 5], the worsening of epidemics of major respiratory viruses (e.g., Influenza, RSV) [21], as well as the incidence of tuberculosis [1, 56, 37, 2]. In the context of the COVID-19 pandemic, it has been shown that the virus spreads faster in larger U.S. cities, pointing to a need for faster responses to contain novel pathogens in larger urban areas [57]. Similarly, in Brazil, the initial growth rates of cases and deaths were found to be higher in large cities, and larger cities were associated with a higher incidence of cases and deaths per capita; however, these patterns varied throughout the course of the pandemic [50]. In the case of STDs, the rates of infection are, on average, higher in urban areas as compared to the national rates in the U.S. [47].

In trying to better understand the underpinnings of these patterns, socioeconomic factors such as education, healthcare accessibility, income, and social inequalities have been identified as relevant to the spread of STDs, although *causal* links are difficult to establish [16, 30, 17, 15, 55, 47]. Moreover, urban centers tend to provide more opportunities for individuals to meet and engage in sexual activities more frequently (see [13] for a discussion), especially in men who have sex with men (MSM) [14]. Moreover, urban centers typically attract younger and more sexually active individuals [7, 49], who may often have lower access to healthcare services (as compared to older and wealthier individuals), especially marginalized groups. This lack of access to healthcare services can, in turn, increase the period in which a sexually active individual remains infectious with an STD (due to lack of awareness of infection status or lack of treatment) [46].

There exists a body of literature studying the mechanistic dynamics of infectious diseases in cities, but studies tend to focus on a single disease in a single city [1, 38, 20]. Thus,

these approaches have limitations in terms of the ability to extract generalizable insights, as the models used are often context-specific and too mathematically complex.

In this work, we propose some modeling approaches and ideas that seek to generalize our understanding of STDs in cities. For this, we build upon previous research in which we have documented broad empirical regularities of sexually transmitted infections in urban settings. We then present novel empirical observations, and then we develop a couple of analytical arguments that, we argue, yield a generalizable understanding of the impact of urbanization on the transmission of STDs.

1.1. Previous empirical findings

Focusing on confirmed case data for chlamydia, gonorrhea, and syphilis from 2007 to 2011 [42], we previously reported some interesting statistical patterns at a *systemic* level [47] by performing a cross-sectional analysis of disease incidence in all Metropolitan Statistical Areas (MSAs) of the U.S. We found that the per capita incidence of these STDs increases systematically with the population size of MSAs (see Figure 1), even after controlling for important socio-economic covariates. Moreover, we identified significant differences in the scaling patterns of these three STDs.

Similar findings were reported regarding infectious diseases such as HIV, Influenza, and Meningitis in Brazil [51]. However, the main drivers of these systemic infection patterns are still not well understood. Identifying these drivers is important, for instance, to help us understand to what extent the high incidence of an infectious disease in a city is due to factors that can be influenced through public health interventions.

2. Deconstructing the scaling patterns

To start teasing out these drivers, we pose the question: *do the scaling features of STDs depend on what "type" of MSAs we focus on?* To address this question, we will investigate in more detail how the associations of urban population size and the transmission dynamics of infectious diseases depend on the properties of the disease and on the socio-economic context of the populations in which they spread.

To that end, here we will introduce an extension of the classical statistical framework for urban scaling in [5, 47]. Briefly, the original framework posits that Y, an urban metric, and N, the population size of an urban area, satisfy the scaling relationship:

$$\mathbb{E}[Y|N_i] = Y_0 N_i^{\alpha}. \tag{1}$$

The subscript indexes a city or urban population, Y_0 is a baseline value for Y, and the scaling exponent α measures the average relative change in Y with respect to N. The

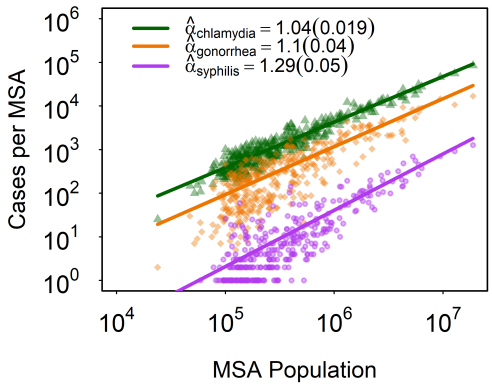

Figure 1: Scaling of chlamydia, gonorrhea, and syphilis incidence with MSA population. Regression lines are based on the Negative Binomial model (2). The $\hat{\alpha}$ estimates represent the slopes of the regressed lines. An estimate greater than 1 suggests a *superscaling* pattern.

scaling relationship in Equation (1) can be transformed to:

$$\log(\mathbb{E}[Y|N_i]) = \log(Y_0) + \alpha \log(N_i), \tag{2}$$

from which empirical regularities can be assessed by estimating the scaling exponent α and the baseline prevalence, Y_0. Using Equation (1), it is easy to show that when $\alpha > 1$ (*superlinear*), the expected incidence rate given by $Y/N = Y_0 N^{\alpha-1}$ is an increasing function of N. In other words, when the scaling relation is *superlinear*, infection *rates* (or per capita incidence) increase with population size.

Herein, we extend the original scaling model in (1) to better understand how socioeconomic factors can impact the scaling patterns for each disease (i.e., as effect modifiers of α and Y_0). Similar to our analysis in [47], the modified model is given by:

$$\mathbb{E}[Y|N_i] = Y_{0j} e^{\xi_j X_{ij}} N_i^{\alpha_{0j} + \alpha_{1j} X_{ij}} \tag{3}$$

where, as in (1), Y is the incidence of cases in a given MSA with population size N_i, X_{ij} is the value of the covariate j in MSA i, with corresponding coefficients α_{0j} and α_{1j} (for the exponent), and ξ_j (for the intercept). The six factors (or covariates) we explore are: *% African American*, *Gini index*, *% Poor*, *education index*, *income per capita*, and *% insured*

(see more details in [47]). The choice of these particular socio-economic covariates reflects what previous studies have found to be key factors in the spread of STDs [30, 17, 55, 39].

In contrast to [47], the modifications in the scaling Equation (3) affect both the intercept, $Y_0^{all} = Y_{0j} e^{\xi_j X_{ij}}$, and the scaling exponent $\alpha^{all} = \alpha_{0j} + \alpha_{1j} X_{ij}$. To see this more clearly, taking logarithm on both sides of (3) yields

$$\log(\mathbb{E}[Y|N_i]) = \underbrace{\left(\log(Y_{0j}) + \xi_j X_{ij}\right)}_{Y_0^{all}} + \underbrace{(\alpha_{0j} + \alpha_{1j} X_{ij})}_{\alpha^{all}} \log(N_i) + \epsilon_i. \tag{4}$$

Given that we are dealing with an overdispersed count variable, we used Negative Binomial regression [31] for models (2) and (4) (see [47] for details).

2.1. New insights based on the extended model

Before delving into the results from the extended model, the six covariates explored here can be broadly categorized based on their correlation with STD per capita rates (see Table 1), as *positively* correlated (*% African American*, *Gini index*, and *% Poor*) and *negatively* correlated (*Education index*, *Income per capita*, and *% Insured*). For example, a city with a larger proportion of low-income residents is expected, on average, to suffer from a larger prevalence of certain STDs, whereas a city with a higher proportion of insured residents is, on average, better equipped to contain the spread of STDs. Note that these observations are based on correlations rather than on *causal* relations.

Covariates	Chlamydia	Gonorrhea	Syphilis
% African American	0.74	0.85	0.58
Gini index	0.25	0.28	0.31
% Poor	0.37	0.31	0.17
Education index	-0.16	-0.12	-0.06
Income per capita	-0.20	-0.17	0.01
% Insured	-0.15	-0.10	-0.18

Table 1: *Pearson correlation coefficients corresponding to each covariate with per capita incidence (incidence/population) across cities. The first three covariates are positively correlated with per capita incidence rates, whereas the last three are negatively correlated with per capita incidence rates (except for the case of income and syphilis, where the correlation is positive but very low).*

Table 2 shows the results of the extended model in (4) that accounts for interaction terms for each of the covariates separately (i.e., one covariate at a time). Of particular interest are the estimates for α_1 and ξ, which represent the effect that each covariate has, respectively, on the baseline prevalence and scaling exponent of the scaling relationships of each STD.

Table 2 indicates that, overall, the *positively* correlated covariates are found to reduce the scaling exponent (i.e., negative $\widehat{\alpha_1}$) while increasing the baseline prevalence (i.e., positive $\widehat{\xi}$), whereas the *negatively* correlated covariates are found to increase the scaling exponent and decrease the baseline prevalence.

While it is a truism to state that different epidemiological contexts affect disease outcomes, these results indicate *how* exactly this occurs through the lens of urban scaling. In brief, these results indicate that in environments that are more auspicious for STD transmission, the scaling exponents tend to be smaller, and intercepts tend to be larger. For the covariate % Poor, for example, this pattern can be interpreted as: i) all else equal, the relative effect of population size on the number of STD cases is lower in MSAs with a high percentage of low-income residents, and ii) all else equal, MSAs with a high percentage of poor residents have larger baseline rates of STDs. That larger levels of the positively correlated covariates lead to higher levels of baseline prevalence is not surprising; however, that larger baseline levels are also associated with lower scaling exponents is not necessarily an intuitive result, nor is it one that can be fully or mechanistically understood using this statistical modeling approach.

Disease	*Covariates*	$\widehat{\alpha_1}$ (p)	$\widehat{\xi}$ (p)
Chlamydia	% African American	-0.004 (0.003)	0.0334 (< 0.001)
	Gini	-1.647 (0.053)	10.758 (0.02)
	% Poor	-0.005 (0.377)	0.045 (0.138)
	Education	-0.0143 (0.549)	0.034 (0.789)
	Income	3.174e-06 (0.408)	-3.134e-05 (0.148)
	% Insured	0.006 (0.085)	-0.039 (0.044)
Gonorrhea	% African American	-0.011 (< 0.001)	0.088 (< 0.001)
	Gini	-2.759 (0.123)	19.062 (0.05)
	% Poor	-0.019 (0.12)	0.141 (0.032)
	Education	0.025 (0.623)	-0.216 (0.426)
	Income	1.110e-05 (0.169)	-9.311e-05 (0.041)
	% Insured	0.010 (0.18)	-0.065 (0.119)
Syphilis	% African American	-0.018 (< 0.001)	0.126 (< 0.001)
	Gini	-5.551 (0.012)	40.133 (0.001)
	% Poor	-0.052 (0.001)	0.340 (< 0.001)
	Education	0.071 (0.278)	-0.516 (0.148)
	Income	2.642e-05 (0.012)	-0.0002 (0.004)
	% Insured	0.023 (0.01)	-0.164 (0.001)

Table 2: *Negative Binomial regression results for all STDs using model* (4)*. In parentheses are the p-values testing the null that* $\alpha_i = 0$ *and* $\xi = 0$*.*

In summary, we found systematic patterns in the incidence of STDs in cities that go beyond a scaling law with universal parameters. Particularly, scaling patterns of disease incidence differ in peculiar ways: factors simultaneously affect the intercept and the exponent such that if one increases, the other decreases, and vice versa. The scaling parameters also vary significantly when focusing on different types of MSAs (i.e., different socioeconomic environments). The fact that STD incidence is well described by scaling laws that change depending on how and what we look at, is not only interesting but can itself indicate how factors interact in cities to facilitate contagion processes.

3. Explaining the scaling patterns

In the previous section, we used a statistical model to better understand how the scaling patterns of STDs in cities change based on different socioeconomic factors. In this section, we will offer a mechanistic understanding of these phenomena to explain the statistical patterns unveiled above and to answer questions such as: *why do the scaling patterns differ between diseases (as shown in [47, 23])?* Or, *why do the scaling exponents for the same disease can have contrasting results in different countries (as shown in [51]) or in different types of MSAs (as we will show below)?* These differences strongly suggest that the intrinsic properties of the diseases and the socioeconomic landscape in which they spread may be at play.

The superlinear scaling of the connectivity of the human contact networks with city size has been suggested as a mechanistic explanation for the superlinear patterns of infections across cities [6, 51]. Indeed, empirical evidence from large mobile phone [54] and Twitter [58] datasets show that social connectivity scales superlinearly with population size. However, it remains unclear how variations in the population size across cities, in combination with the non-linear effects of large social networks, can impact the scaling properties of epidemic outcomes.

To better make sense of the differences in the scaling patterns of the STDs described above, we can focus on their intrinsic transmission capacity and assess how their symptomatology and/or inherent infectiousness are key factors in the overall transmissibility of these pathogens.

Chlamydia, having a high infection risk per sexual act and being relatively asymptomatic (i.e., difficult to detect and prevent), features an advantageous combination of traits that makes it the most transmissible of the three STDs studied herein. Syphilis, on the opposite end, has lower infectiousness and is the most symptomatic of the three STDs [43]. As a result, chlamydia is accountable for a larger number of infections as compared to syphilis

which, presumably, propagates largely through high-risk sexual networks [28] where individuals engage in riskier sexual practices.

Hence, it is not surprising that in any given city, the prevalence and incidence of chlamydia are higher than that of syphilis. But how sensitive are these two diseases to increases in population size? Figure 1 shows that the associations of incidence with population size (i.e., the slope of the lines) decrease as STDs increase their transmissibility and prevalence. Therefore, the scaling analysis indicates that population size is more strongly associated with the spreading capacity of STDs that transmit less readily. According to this view, the rate of contagion of pathogens that transmit with less ease, like syphilis, is heavily dependent on the characteristics of the environment through which they can spread.

In light of this, a thread of plausible mechanisms to explain the observed scaling patterns emerges. First, the rate of spread of an STD that transmit less easily, such as syphilis, should be highly sensitive to the connectivity patterns in a city, which is, in turn, a function of population size; thus, the steeper increase of incidence with population size and, consequently, the larger scaling exponent for an STD like syphilis. Second, STDs that transmit less easily are less prevalent, thus the negative correlation between scaling exponents and intercepts. And third, we expect that prevalence levels of less transmissible STDs will be subject to higher levels of variation due to larger stochastic effects on the spreading dynamics of the disease (especially in smaller cities).

Consequently, we hypothesize that, in general, the lower the spreading capacity of a disease due to its inherent transmissibility or the environment in which it spreads, i) the more its transmission becomes contingent on population size, ii) the lower its overall incidence, and iii) the larger its variation in incidence. To test the validity and generalizability of these insights, we will turn to, first, mathematical models of infection and, second, a new theory of urban scaling.

3.1. Explaining the scaling patterns using a simple epidemiological model

The SIS (Susceptible-Infected-Susceptible) epidemic model is widely used to study the dynamics of sexually transmitted diseases in homogeneously [33] and heterogeneously mixed populations [9, 35, 8]. It is a simplified representation of an endemic disease with no lasting immunity. Therefore, this modeling approach, albeit simple, can be suitable for explaining some of the scaling patterns observed in the data of the three bacterial STDs studied above.

In this framework, individuals are classified based on their infectious status as susceptible (S) or infected (I). Individuals become infected at rate β and recover at rate γ, where $1/\gamma$ is the expected time to recover. The transmission rate is defined as the per capita contact rate, c, multiplied by the probability a contact with an infectious individual leads to an infection, p; that is, $\beta = c \times p$. Intuitively, it is clear that the number of contacts

experienced by a random individual in a population depends on the size of the population, $N = S + I$, and presumably, c should increase with population size [52].

Without making any assumptions about the functional form of $\beta(N) = c(N) \times p$, and considering a relatively short modeling time frame (so that we can neglect demographic aspects such as births and deaths), the system describing the disease dynamics is:

$$\frac{dS}{dt} = -\beta(N) S I/N + \gamma I \tag{5}$$

$$\frac{dI}{dt} = \beta(N) S I/N - \gamma I. \tag{6}$$

This system has a basic reproductive number [10] given by $R_0(N) = \beta(N)/\gamma$, and the endemic equilibrium can be found by solving $dI/dt = 0$, yielding

$$I^*(N) = N[1 - 1/R_0(N)]. \tag{7}$$

In the simplest case, we can assume that c is constant, and therefore $R_0(N) = R_0$ (often suitable when $N \to \infty$). In this scenario, $I^*(N)$ is a linear function of N; hence that assumption does not lead to superlinear scaling.

Building upon the empirical evidence that indicates that social connectivity scales superlinearly with population size [54, 58], we can instead assume that the average number of contacts per individual, $\langle c \rangle$ also scales with population size as $\langle c \rangle \sim N^{\psi}, 0 < \psi < 1$, and the empirical estimations of ψ fluctuate around 0.15 (0.11 to 0.21) [58] depending on the level of aggregation and country. Thus, we postulate that the contact rate, $c(N)$, is a function of population size given by:

$$c(N) = bN^a, \tag{8}$$

where a and b are positive constants. The expression of the transmission rate, $\beta(N) = c(N) \times p$, has one component which is affected by population size, i.e., $c(N)$ and p, which depends on the intrinsic transmissibility of the pathogen but also on sexual behaviors (e.g., use of condoms)[1].

[1]Noteworthy, the baseline parameter b can be a function of socioeconomic conditions. For example, cities with a large fraction of high-risk individuals (i.e., individuals with a large number of sexual partners) would feature a larger b. Since the mode of transmission of the disease (e.g., sexual, airborne) is crucial in defining what constitutes a contact, this can also affect the value of b. Additionally, we note that the

Substituting Equation (8) in (7) yields:

$$I^*(N) = N\left[1 - \frac{1}{\frac{pb}{\gamma}N^a}\right]. \tag{9}$$

To understand the impact of the parameters p, b, and γ on the scaling of disease prevalence with population size[2], we first need to rewrite Equation (9) in terms of $\log(I^*)$ and $\log(N)$. Taking logs on both sides of (9), and after some transformations, we obtain:

$$\log(I^*) = \log(N) + \log\left(1 - \frac{e^{-a\log(N)}}{\frac{pb}{\gamma}}\right). \tag{10}$$

Taking partial derivatives in (10) results in:

$$\frac{\partial \log(I^*)}{\partial \log(N)} = 1 + \frac{a}{\frac{pb}{\gamma}e^{a\log(N)} - 1} = 1 + \frac{a}{R_0(N) - 1}. \tag{11}$$

When the disease can spread, that is, when $R_0 > 1$, the denominator in (11) is positive, and so the derivative is larger than 1, which is indicative of the superscaling of incidence. To see this, note that based on Equation (2), the derivative $\partial \log(Y)/\partial \log(N) = \alpha$, which in combination with (11) renders the following relation:

$$\alpha = 1 + \frac{a}{R_0(N) - 1}. \tag{12}$$

From this expression, we see that when $a = 0$ (i.e., the contact rate does not increase with population size), the scaling exponent becomes 1 (linear). Equation (12) allows us to determine the effect of the parameters on the scaling properties of disease prevalence. As $R_0(N) = \frac{pb}{\gamma}N^a$ increases, the effect of population size on the scaling of prevalence is reduced. In other words, if the disease becomes more transmissible–either due to becoming more infectious (larger p), or due to becoming harder to detect and treat (larger $1/\gamma$), or

socioeconomic context can affect not only the infection risk p (e.g., via better sex education and easy access to condoms), but it can also affect the recovery rate parameter γ. For instance, higher income is associated with better access to health care and treatment, both of which are critical in determining the speed of recovery. Thus, improved socioeconomic conditions can reduce $1/\gamma$ [46].

[2]If we assume that the system under study has reached its equilibrium, then the cumulative incidence from time τ_1 to τ_2 is equal to prevalence I^* times a constant. More specifically, $Incidence(\tau_1, \tau_2) = \gamma(\tau_1 - \tau_1)I^*$. Hence, at equilibrium, the scaling properties of prevalence and incidence are equivalent.

due to sexual behaviors changing as to increase the rate of sexual contacts in the population (larger b)–then the effect of population size in the contagion process is diminished.

Moreover, the observation that larger scaling exponents correlate with lower intercepts (i.e., prevalence) can also be explained by differences in transmissibility of the STDs (p) combined with increased connectivity in larger populations. That is, as R_0 increases, α decreases, while the prevalence increases (see Equation (7)).

To visualize these findings, Figure 2 shows the plot of three curves corresponding to three diseases with different infectiousness p. From the figure, we can readily observe that the effect of population size on prevalence decreases as p increases, while the prevalence levels increase with p.

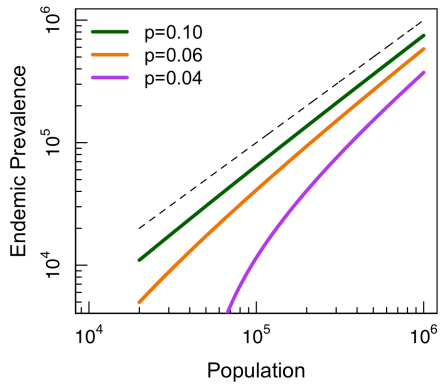

Figure 2: Disease prevalence versus population size for different intrinsic transmissibility values (p). As p increases, the marginal effect of the population size decreases (i.e., lower scaling exponent). Other parameters: $a = 0.15, b = 0.5, \gamma = 1/10$. Note the similarity with the left panel in Figure 1.

Finally, note that Equation (12) indicates that the derivative $\frac{\partial \log(I^*)}{\partial \log(N)}$ depends on N, unlike the case in (2) where α, by design, is not a function of N. However, as Figure 5 in the Appendix shows, the relationship between the log incidence and log population is only linear for relatively high populations, with a non-linear decaying for lower populations. Remarkably, this is precisely the behavior we obtain using a simple SIS model. Moreover, this behavior at low population sizes implies that changes in the environment can have a relatively large impact on small populations.

A key feature of Figure 1 is the substantial variability across cities of similar size, which cannot be obtained from the deterministic model. To recreate this variability, in the Appendix, we conducted simulations using the stochastic version of the SIS model presented above. Indeed, the SIS stochastic simulations provide scaling patterns more similar to that of the empirical data in Figure 1 than the deterministic SIS model in Figure 2.

3.2. Explaining the scaling patterns through the lens of a recent theory of urban scaling

In light of the scaling patterns observed for the incidence of STDs, as well as for a wide and diverse set of urban phenomena (e.g., education, employment, innovation, and crime), we recently introduced a novel theory of urban scaling to explain the differences in, and relationships between, prevalence, scaling exponents, and cross-sectional variance of all these urban phenomena [23, 22]. This was achieved by coupling ideas from the fields of economic complexity (i.e., any given social phenomenon can only occur if a number of different but complementary factors are simultaneously present) [29, 25], and cultural evolution (i.e., the diversity of factors grows logarithmically with population size) [26, 27]. The resulting model predicts that: compared to a less complex phenomenon (i.e., one that requires fewer complementary factors to occur), a more complex phenomenon is expected to be less prevalent (rarer), scale more steeply with population size (more super-linearly), and show larger variance across cities of similar size.

In the Appendix, we show that the complexity of a phenomenon is determined by three aspects: 1) its *inherent* difficulty (e.g., pathogen transmissibility in the context of STDs), 2) the social fabric (e.g., a network of interactions between individuals) in which the phenomenon occurs, and 3) the capacity of the individuals involved to contribute to the occurrence of the phenomenon (e.g., behavioral practices to prevent or enhance the spread of infections).

Here, we illustrate the concepts and the link of the new scaling theory to the empirical evidence discussed above regarding syphilis and chlamydia. Figure 3, left panel, displays the three aspects described above: compared to syphilis, chlamydia has a higher prevalence, lower scaling exponents, and lower variability. As before, we hypothesize that a key driver of the diverging scaling patterns of these two STDs is the difference in their inherent transmissibility. Chlamydia is more infectious and less symptomatic than syphilis, hence the spreading capacity of chlamydia is considerably higher than that of syphilis. This observation is consistent with our theory once we make the connection between complexity and (lack of) transmissibility. Syphilis is more complex than chlamydia in the sense that syphilis requires the occurrence of more factors to be transmitted than chlamydia does. For example, given its lower infectiousness per sexual contact compared to chlamydia, syphilis requires, on average, more sexual interactions for a transmission event to occur. Also,

since it is more symptomatic to be transmitted, people may have to engage in risky sexual behavior despite the presence of symptoms.

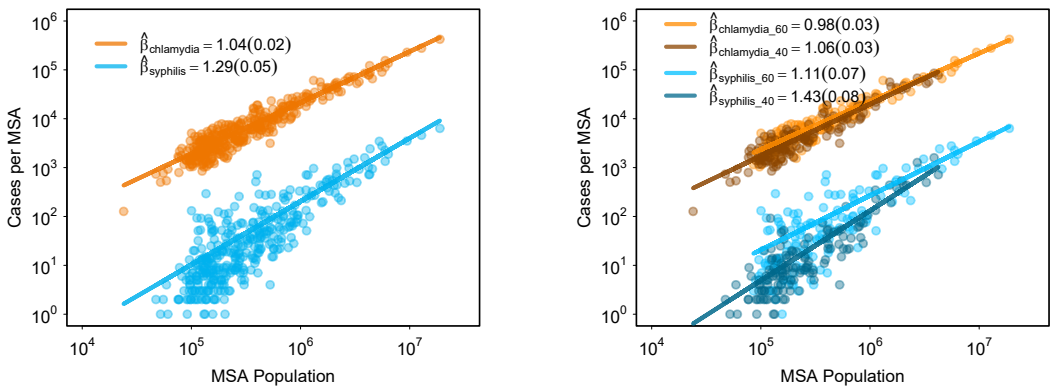

Figure 3: (left) Scaling patterns of the 5-year average incidence of STDs stratified by disease (chlamydia [orange] and syphilis [blue]) in metropolitan statistical areas (MSA) in the U.S. (right) Scaling patterns of STDs stratified by disease and by gini index of cities (cities with high-income inequality [light colors] and cities with low-income inequality [dark colors]). The two groups of cities are defined, respectively, as those that belong to the two highest quintiles or the top 40th percentile (light blue) and the two lowest quintiles or the bottom 40th percentile (dark blue) in terms of the Gini coefficient. The legends show the corresponding scaling exponents and their standard errors.

The right panel of Figure 3 shows how the scaling properties of syphilis and chlamydia are not only disease-dependent but also context-dependent (note that Figure 3 represents a visual confirmation of the observations derived from the results in Table 2). In this example, the "context" is determined by the level of income inequality measured by the Gini coefficient. Focusing on the case of syphilis, the prevalence of syphilis is higher while its scaling exponent and variability are lower in cities with higher levels of income inequality (i.e., cities belonging to the two highest quintiles with respect to their Gini coefficients) than in cities with lower levels of income inequality (i.e., cities belonging to the two lowest quintiles). Similar divergent patterns are observed if cities are instead categorized by the other covariates presented in the previous section. All of these metrics are a reflection of socioeconomic aspects that affect the properties of the populations and the sexual networks on which the disease is spreading. The right panel of Figure 3 shows that the incidence of syphilis is higher, with a lower scaling exponent and variability, in an environment that is presumably more auspicious for the spread of STDs.

In sum, the left panel of Figure 3 shows how, everything else constant (e.g., the environment), a difference in the inherent complexity of two phenomena (e.g., pathogen transmissibility) affects, in very specific ways, how these STDs scale and vary with population size. The right panel demonstrates that controlling for the inherent complexity of the STD, the propensity of the environment (e.g., level of income inequality) also affects the scaling patterns.

3.2.1 Complexity and predictability of phenomena

The findings of the current study are interestingly connected to the insights from a previous study on the predictability of infectious diseases [53], which showed that the time series of diseases with lower R_0 (i.e., lower infection potential) are, on average, harder to predict. Moreover, diseases that spread in networks with topologies that are less conducive to disease spread are also less predictable. These findings make sense intuitively, given that *rarer* events are typically more difficult to predict.

To elucidate the connection, we need to note three other links. First, predictability, as defined in [53] (using permutation entropy as a model-independent measure of predictability), is related to Kolmogorov complexity [48], which is the length of the shortest computer program that produces the object (phenomenon) in question. Second, in our new scaling theory, more complex phenomena are, by definition, those requiring more elements to occur. Lastly, more complex phenomena feature more *cross-sectional* variability, and consequently, the occurrence of more complex phenomena is more difficult to predict, which closes the explanatory loop with the findings in [53]. It is also interesting to note how the cross-sectional variability of disease spread is lower both when diseases have lower infection capacity (intrinsic property of the phenomenon) and when diseases spread in less conducive environments (e.g., social networks), which is nicely tied into how we decomposed the concept of complexity.

The intrinsic complexity of a disease affects not only the *cross-sectional* (i.e., in cities of a given population size) predictability of the disease spreading process but also its predictability in *time* (i.e., a given disease in a city). To show this, we use permutation entropy, akin to [53], to measure the predictability (defined as 1-permutation entropy) of infection time series for the three STDs in the dataset.

As a complementary way to measure the variability of the time series of the three diseases across cities, we also computed the coefficient of variation (COV=standard deviation/mean) of each disease in each city.

Figure 4 shows the distributions of predictabilities (left panel) and COV (right panel) for each disease across cities. These figures show that for diseases with lower infection capacity (e.g., syphilis), the time series predictability of the spreading process is lower, and

the variability is larger, in line with the findings in [53] and with what we noted regarding cross-sectional variability.

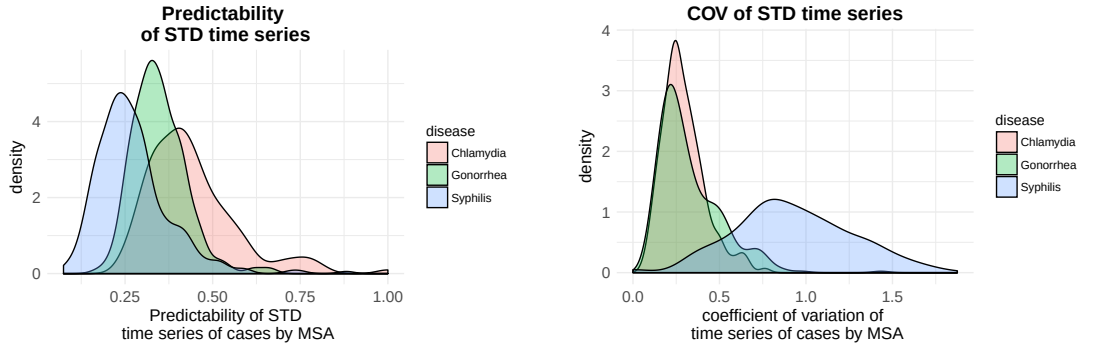

Figure 4: Distributions of predictabilities (1-permutation entropies) (left panel) and COV (right panel) for each disease across cities. We compute the permutation entropy of the time series of each disease in each city with at least 10 years of annual incidence data (to ensure a more robust/data-rich measure of permutation entropy). The chlamydia, gonorrhea, and syphilis time series data span from 1996, 1995, and 1984, respectively, until 2011. We used the 'statcomp' package in R and a sliding window size of 4 (we also tried window sizes of 3 and 5, and the results were similar).

4. Discussion

The first contribution of this study was to present empirical evidence demonstrating that the statistical patterns of STD incidence in cities are systematically shaped by the intrinsic and contextual factors that determine the spreading capacity of the pathogen. Specifically, the three more salient features of the scaling patterns, namely the scaling exponent, the baseline prevalence, and the cross-sectional variance of the incidence depend in a predictable way on the inherent infectiousness of the infectious diseases and the socioeconomic properties of the cities in which they spread. Diseases that spread more easily are associated with lower scaling exponent, higher prevalence, and lower variance.

Going beyond the statistical description of these empirical relationships, the second contribution was to present two mechanistic models that provided compelling explanations for these patterns and their underlying drivers. In the first approach, we used a simple model of infectious disease spread (SIS model) that incorporated in its mathematical formulation the empirical observation that contact rates increase with population size. This model, in its deterministic and stochastic versions, helped us confirm that the scaling patterns of diseases with higher transmission rates have lower scaling exponent, higher prevalence, and

lower variance. Additionally, this model also captured the non-linear (in the log-log scale) decay of the incidence with population size for cities at the lower end of the population size spectrum. This behavior is not easily obtained with traditional scaling models, but it emerged remarkably naturally from this simple model.

In the second approach, we used a recently published theory of urban scaling to provide coherent and intuitive explanations of these empirical patterns. Specifically, the theory had previously revealed an insightful conceptual dissection of the scales at which the different facets of the "complexity" of a phenomenon operate. The theory suggests that the *net complexity* of a phenomenon results from the multiplicative effects of the inherent complexity of the phenomenon, the level of engagement of the individuals involved, and the conduciveness of the surrounding environment for the phenomenon to occur. Translated to the context of infectious diseases, these three aspects can be intuitively interpreted as the intrinsic infectiousness of the disease, the susceptibility of individuals, and the suitability for disease spread of the underlying social fabric, respectively. These factors act at different levels (i.e., intra-individual, inter-individual, and community levels) and are interrelated. The first factor is tied to the biology of the pathogen and the host, with the risk of infection per sexual act also varying with the mode of transmission and the symptomatology of the disease. The second term relates to individuals' risk behavior (e.g., condom use, number of sexual partners) and perceived risk, which could be, in turn, affected by educational level, and access to medical and counseling services. Lastly, the third term refers to aspects such as the sexual network in which individuals are embedded and the prevalence of the disease in the community, which is, in turn, governed by social and cultural norms and the socioeconomic landscape of the community.

Inspired by this conceptual framework and the recent work in [53], a third contribution was to show that the intrinsic complexity of a disease not only determines the cross-sectional variability of disease incidence but also its temporal predictability.

The two modeling approaches introduced herein to explain the scaling patterns of STDs in cities are based on different premises and assumptions. The contagion model is a dynamic non-linear formulation of a spreading process in time, whereas the scaling theory is based on probabilistic formulations based on first principles and does not explicitly incorporate time. However, there are a number of interesting parallels between the two approaches. For example, R_0 and *net complexity* are inversely related. We showed how both these metrics provide a convenient mathematical structure that lends itself to helpful conceptual descriptions of the drivers of these contagion processes.

These two modeling approaches suggest that the factors that influence the urban scaling patterns of STDs can be grouped into three categories: 1) those related to the suitability of the environment (level of population density and mobility, quality of public health in-

frastructure, presence of inequalities in healthcare access), 2) those related to the specific characteristics of the disease (e.g., symptom profile, inherent transmissibility), and 3) those related to the behaviors of urban residents (e.g., frequency of interactions that can lead to disease transmission).

The connectivity of sexual networks [36], in combination with their size, can influence the transmissibility of infectious diseases [60, 40, 45, 44, 32]. Arguably, the structure and size of high-risk sexual networks are affected by the size of the cities in which they exist. For example, larger cities can foster either larger or more numerous, high-risk sexual networks. In other words, the higher rates of STD prevalence in large cities could be explained, at least in part, by a disproportionate presence of high-risk groups in those populations. Therefore, one interesting avenue to improve the contagion model used here is to relax the homogenous mixing assumption and instead assume that individuals are embedded in a network of (sexual) interactions. In this setting, it would be interesting to explore how the network structure (e.g., degree distribution) affects the scaling patterns. In fact, based on the current findings, we hypothesize that populations featuring network topologies that are less suitable for the widespread of infections (e.g., networks with low contact heterogeneity) will lead to scaling patterns of incidence characterized by high scaling exponents, low baseline prevalence, and high cross-sectional variability. Additionally, the variability in the spreading process due to the network structure could further explain the amount of variance seen in the real-world data that was not captured in the stochastic model presented herein. Within a network framework, it would also be possible to investigate the effects of different modes of transmission (e.g., sexually transmitted or direct contact versus airborne) in the scaling patterns.

The insights revealed in this study can be translated into policy recommendations. The scaling analysis indicated that STDs that transmit with less ease feature a larger superlinear effect. Therefore, resources and efforts to contain the spread of less prevalent diseases should focus on larger cities. Interestingly, a disease containment effort that distributes resources uniformly across all populations, instead of focusing on larger cities, may unintendedly lead to lower overall disease prevalence but simultaneously lead to a scenario in which larger cities are even more disproportionately affected by the disease than before the containment effort came into effect. Another critical insight is that a contagion-favorable socioeconomic landscape has an effect on the scaling patterns akin to the one that the intrinsic infectivity of a disease has. Therefore, if the scaling patterns of a given infectious disease in two sets of cities (e.g., cities within two geographic regions of the U.S. or cities within two countries) differ substantially, we could hypothesize that the cities in which the scaling pattern had a lower exponent and higher prevalence have, on average, a more conducive environment for the spread of the disease. Also, in [22], we showed that all else equal, changes in the environment's suitability for a phenomenon to occur lead to larger

changes in outcomes than comparable changes in individuals' behaviors. In the context of STDs, this observation would imply that interventions that decrease the suitability of the environment for STD transmission (e.g., providing better healthcare infrastructure and resources, or better epidemiological monitoring of neighborhoods or establishments with a high frequency of risky sexual encounters) are generally more impactful than interventions directed towards changing the behavior of individuals. This type of information may be useful to a policymaker that is, for instance, trying to determine how to effectively deploy limited resources into two or more countries/regions to contain the spread of a particular infectious disease.

Acknowledgments

We would like to thank Julia Wei Wu for the valuable discussions that helped shape some sections of this study.

References

[1] Emilie Alirol et al. "Urbanisation and infectious diseases in a globalised world". In: *The Lancet infectious diseases* 11.2 (2011), pp. 131–141. DOI: 10.1016/S1473-3099(10)70223-1.

[2] Juan Pablo Aparicio, Angel Capurro, and Carlos Castillo-Chávez. "Markers of disease evolution: the case of tuberculosis". In: *Journal of Theoretical Biology* 215.2 (2002), pp. 227–237. DOI: 10.1006/jtbi.2001.2489.

[3] Juan Pablo Aparicio and Carlos Castillo-Chavez. "Mathematical modelling of tuberculosis epidemics". In: *Math Biosci Eng* 6 (2009), pp. 209–37. DOI: 10.3934/mbe.2009.6.209.

[4] Pwaveno Bamaiyi. "The Role of Demographics and Human Activities in the Spread of Diseases". In: *Current Trends in Technology and Sciences* 2 (Jan. 2013), pp. 253–257.

[5] Luís Bettencourt et al. "Growth, innovation, scaling, and the pace of life in cities". In: *Proceedings of the National Academy of Sciences* 104.17 (2007), pp. 7301–7306. DOI: 10.1073/pnas.0610172104.

[6] Luís MA Bettencourt and Geoffrey B West. "Bigger cities do more with less". In: *Scientific American* 305.3 (2011), pp. 52–53. DOI: 10.1038/scientificamerican0911-52.

[7] Diane Binson et al. "Prevalence and social distribution of men who have sex with men: United States and its urban centers". In: *Journal of Sex Research* 32.3 (1995), pp. 245–254. DOI: 10.1080/00224499509551795.

[8] Carlos Castillo-Chavez, Wenzhang Huang, and Jia Li. "Competitive exclusion and coexistence of multiple strains in an SIS STD model". In: *SIAM Journal on Applied Mathematics* 59.5 (1999), pp. 1790–1811. DOI: 10.1137/s0036139997325862.

[9] Carlos Castillo-Chavez, Wenzhang Huang, and Jia Li. "The effects of females susceptibility on the coexistence of multiple pathogen strains of sexually transmitted diseases". In: *Journal of Mathematical Biology* 35.5 (1997), pp. 503–522. DOI: 10.1007/s002850050063.

[10] Gerardo Chowell and Fred Brauer. "The Basic Reproduction Number of Infectious Diseases: Computation and Estimation Using Compartmental Epidemic Models". In: *Mathematical and Statistical Estimation Approaches in Epidemiology*. Ed. by Gerardo Chowell et al. Dordrecht: Springer Netherlands, 2009, pp. 1–30. DOI: 10.1007/978-90-481-2313-1_1.

[11] David P Clark. *Germs, Genes, & Civilization: How Epidemics Shaped who We are Today*. FT Press, 2010.

[12] C Cohen, D Horlacher, and F Landis MacKellar. "Is urbanization good for a nation's health". In: European Population Conference. Vienna, Austria, Sept. 2010.

[13] Alan Collins. "Sexuality and sexual services in the urban economy and socialscape: an overview". In: *Urban Studies* 41.9 (2004), pp. 1631–1641. DOI: 10/dpvxm4.

[14] Alan Collins and Stephen Drinkwater. "Fifty shades of gay: Social and technological change, urban deconcentration and niche enterprise". In: *Urban Studies* 54.3 (2017), pp. 765–785. DOI: 10.1177/0042098015623722.

[15] Richard Crosby et al. "Social capital as a predictor of adolescents' sexual risk behavior: a state-level exploratory study". In: *AIDS and Behavior* 7.3 (2003), pp. 245–252. DOI: 10.1023/A:1025439618581.

[16] Ralph DiClemente et al. "Prevention and control of sexually transmitted infections among adolescents: the importance of a socio-ecological perspective: a commentary". In: *Public Health* 119.9 (2005), pp. 825–836. DOI: 10.1016/j.puhe.2004.10.015.

[17] Thomas Farley. "Sexually transmitted diseases in the Southeastern United States: location, race, and social context". In: *Sexually transmitted diseases* 33.7 (2006), S58–S64. DOI: 10.1097/01.olq.0000175378.20009.5a.

[18] Sandro Galea. "Urbanization, urbanicity, and health". In: *Journal of urban health: Bulletin of the New York Academy of Medicine* 79.1 (2002), S1–S12. DOI: 10.1093/jurban/79.suppl_1.s1.

[19] Sandro Galea, Nicholas Freudenberg, and David Vlahov. "Cities and population health". In: *Social science & medicine* 60.5 (2005), pp. 1017–1033. DOI: 10.1016/j.socscimed.2004.06.036.

[20] Dionne Gesink et al. "Sexually transmitted disease core theory: roles of person, place, and time". In: *American journal of epidemiology* 174.1 (2011), pp. 81–89. DOI: 10.1093/aje/kwr035.

[21] W Paul Glezen. "The changing epidemiology of respiratory syncytial virus and influenza: impetus for new control measures". In: *The Pediatric infectious disease journal* 23.11 (2004), S202–S206. DOI: 10.1097/01.inf.0000144662.86396.07.

[22] Andres Gomez-Lievano and Oscar Patterson-Lomba. "Estimating the drivers of urban economic complexity and their connection to economic performance". In: *Royal Society open science* 8.9 (2021), p. 210670. DOI: 10.1098/rsos.210670.

[23] Andres Gomez-Lievano, Oscar Patterson-Lomba, and Ricardo Hausmann. "Explaining the prevalence, scaling and variance of urban phenomena". In: *Nature Human Behaviour* 1 (2016), p. 0012. DOI: 10.1038/s41562-016-0012.

[24] Trudy Harpham and Catherine Molyneux. "Urban health in developing countries: a review". In: *Progress in Development Studies* 1.2 (2001), pp. 113–137. DOI: 10.1177/146499340100100202.

[25] Ricardo Hausmann and César A. Hidalgo. "The network structure of economic ouput". In: *Journal of Economic Growth* 16 (Oct. 2011), pp. 309–342. DOI: 10.1007/s10887-011-9071-4.

[26] Joseph Henrich. "Demography and Cultural Evolution: How Adaptive Cultural Processes can Produce Maladaptive Losses: The Tasmanian Case". In: *American Antiquity* 69.2 (2004), pp. 197–214. ISSN: 00027316. DOI: 10.2307/4128416.

[27] Joseph Henrich. *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press, 2015.

[28] Herbert W Hethcote and James A Yorke. *Gonorrhea transmission dynamics and control*. Vol. 56. Springer Berlin, 1984.

[29] César A. Hidalgo and Ricardo Hausmann. "The building blocks of economic complexity". In: *PNAS* 106.25 (June 2009), pp. 10570–10575. DOI: 10.1073/pnas.0900943106.

[30] David R Holtgrave and Richard A Crosby. "Social capital, poverty, and income inequality as predictors of gonorrhoea, syphilis, chlamydia and AIDS case rates in the United States". In: *Sexually transmitted infections* 79.1 (2003), pp. 62–64. DOI: 10.1136/sti.79.1.62.

[31] Noriszura Ismail and Abdul Aziz Jemain. "Handling overdispersion with negative binomial and generalized Poisson regression models". In: *Casualty Actuarial Society Forum*. 2007, pp. 103–158.

[32] Matt Keeling. "The implications of network structure for epidemic dynamics". In: *Theoretical population biology* 67.1 (2005), pp. 1–8. DOI: 10.1016/j.tpb.2004.08.002.

[33] Matt J. Keeling and Pejman Rohani. *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press, 2008.

[34] David A Leon. "Cities, urbanization and health". In: *International journal of epidemiology* 37.1 (2008), pp. 4–8. DOI: 10.1093/ije/dym271.

[35] Jia Li et al. "Coexistence of pathogens in sexually-transmitted disease models". In: *Journal of Mathematical Biology* 47.6 (2003), pp. 547–568. DOI: 10.1007/s00285-003-0235-5.

[36] Fredrik Liljeros et al. "The web of human sexual contacts". In: *Nature* 411.6840 (2001), pp. 907–908. DOI: 10.1038/35082140.

[37] Knut Lönnroth et al. "Drivers of tuberculosis epidemics: the role of risk factors and social determinants". In: *Social science & medicine* 68.12 (2009), pp. 2240–2246. DOI: 10.1016/j.socscimed.2009.03.041.

[38] Liang Mao and Ling Bian. "Spatial–temporal transmission of influenza and its health risks in an urbanized area". In: *Computers, Environment and Urban Systems* 34.3 (2010), pp. 204–215. DOI: 10.1016/j.compenvurbsys.2010.03.004.

[39] Michael Marmot. "Social determinants of health inequalities". In: *The Lancet* 365.9464 (2005), pp. 1099–1104. DOI: 10.1016/S0140-6736(05)71146-6.

[40] Robert M. May and Alun L. Lloyd. "Infection dynamics on scale-free networks". In: *Phys. Rev. E* 64 (6 Nov. 2001), pp. 66–112. DOI: 10.1103/PhysRevE.64.066112.

[41] David M Morens and Anthony S Fauci. "Emerging Infectious Diseases: Threats to Human Health and Global Stability". In: *PLoS pathogens* 9.7 (2013), e1003467. DOI: 10.1371/journal.ppat.1003467.

[42] National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (U.S.). Division of STD Prevention. *Sexually Transmitted Disease Surveillance 2012*. 2014, pp. 1–174. URL: https://stacks.cdc.gov/view/cdc/21548.

[43] National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (U.S.). Division of STD Prevention. *Sexually transmitted disease surveillance 2019 : National overview - Sexually Transmitted Disease Surveillance, 2019*. 2021. URL: https://stacks.cdc.gov/view/cdc/105137.

[44] Mark E.J. Newman. "Spread of epidemic disease on networks". In: *Phys. Rev. E* 66 (1 July 2002), p. 016128. DOI: 10.1103/PhysRevE.66.016128.

[45] Romualdo Pastor-Satorras and Alessandro Vespignani. "Epidemic dynamics and endemic states in complex networks". In: *Phys. Rev. E* 63 (6 May 2001), p. 066117. DOI: 10.1103/PhysRevE.63.066117.

[46] Oscar Patterson-Lomba et al. "Modeling the role of healthcare access inequalities in epidemic outcomes". In: *Mathematical Biosciences & Engineering* 13.5 (2016), p. 1011. DOI: 10.3934/mbe.2016028.

[47] Oscar Patterson-Lomba et al. "Per capita incidence of sexually transmitted infections increases systematically with urban population size: a cross-sectional study". In: *Sexually transmitted infections* (2015), sextrans–2014. DOI: 10.1136/sextrans-2014-051932.

[48] Antonio Politi. "Quantifying the Dynamical Complexity of Chaotic Time Series". In: *Physical review letters* 118.14 (2017), p. 144101. DOI: 10.1103/PhysRevLett.118.144101.

[49] Thomas Quinn. "Population migration and the spread of types 1 and 2 human immunodeficiency viruses". In: *Proceedings of the National Academy of Sciences* 91.7 (1994), pp. 2407–2414. DOI: 10.1073/pnas.91.7.2407.

[50] Haroldo V Ribeiro et al. "City size and the spreading of COVID-19 in Brazil". In: *PloS one* 15.9 (2020), e0239699. DOI: 10.1371/journal.pone.0239699.

[51] Luis Rocha, Anna Thorson, and Renaud Lambiotte. "The non-linear health consequences of living in larger cities". In: *Journal of urban health: Bulletin of the New York Academy of Medicine* 92.5 (2015). DOI: 10.1007/s11524-015-9976-x.

[52] Lisa Sattenspiel. *The geographic spread of infectious diseases: models and applications*. Princeton University Press, 2009.

[53] Samuel V Scarpino and Giovanni Petri. "On the predictability of infectious disease outbreaks". In: *Nature Communications* 10.1 (2019), p. 898. DOI: 10.1038/s41467-019-08616-0.

[54] Markus Schläpfer et al. "The scaling of human interactions with city size". In: *Journal of The Royal Society* 11 (2014). DOI: 10.1098/rsif.2013.0789.

[55] Salaam Semaan et al. "Social capital and rates of gonorrhea and syphilis in the United States: spatial regression analyses of state-level associations". In: *Social science & medicine* 64.11 (2007), pp. 2324–2341. DOI: 10.1016/j.socscimed.2007.02.023.

[56] Carolyn Stephens. "Healthy cities or unhealthy islands? The health and social implications of urban inequality". In: *Environment and Urbanization* 8.2 (1996), pp. 9–30. DOI: 10.1177/095624789600800211.

[57] Andrew Stier, Marc Berman, and Luís Bettencourt. "Early pandemic COVID-19 case growth rates increase with city size". In: *npj Urban Sustainability* 1.1 (2021), pp. 1–6. DOI: 10.1038/s42949-021-00030-0.

[58] Michele Tizzoni et al. "The scaling of human contacts and epidemic processes in metapopulation networks". In: *Scientific reports* 5.1 (2015), pp. 1–11. DOI: 10.1038/srep15111.

[59] United Nations. *World Urbanization Prospects: The 2011 Revision*. URL: https://desapublications.un.org/publications/world-urbanization-prospects-2011-revision.

[60] Duncan Watts and Steven Strogatz. "Collective dynamics of small-world networks". In: *nature* 393.6684 (1998), pp. 440–442. DOI: 10.1038/30918.

[61] Ping Zhang and Peter M Atkinson. "Modelling the effect of urbanization on the transmission of an infectious disease". In: *Mathematical biosciences* 211.1 (2008), pp. 166–185. DOI: 10.1016/j.mbs.2007.10.007.

5. Appendix

5.1. Results using the classic scaling model

Using the classical model in (2) we obtain $\alpha > 1$ for all three STDs, indicating that the rates of STDs appear to increase in a systematic way with population size (superlinear scaling), as shown in Figure 1 and Table 3.

Covariates	Chlamydia	Gonorrhea	Syphilis
$\widehat{\alpha}$	1.04	1.10	1.29
(SE)	(0.019)	(0.040)	(0.051)
$\log(\widehat{Y_0})$	-1.94	-2.84	-5.46
(SE)	(0.105)	(0.219)	(0.280)
Observations	364	364	364

Table 3: *Negative Binomial regression results in the four infectious diseases using model* (2).

5.2. Loess fit to the data

Figure 5 shows that fitting a smooth function (i.e., loess regression, a non-parametric method that is in principle agnostic to the shape of the data) to the log of the incidence versus the log of the population (instead of a straight line) reveals that the relationship

between of the log incidence and log population is only linear for relatively high populations, with a non-linear decaying for lower populations.

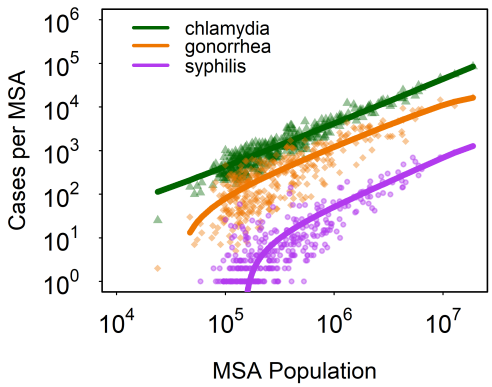

Figure 5: Incidence of chlamydia, gonorrhea, and syphilis versus MSA population, fitted with a loess line. Note the similarity with Figure 2.

5.3. Stochastic SIS model

A key feature of Figure 1 is the substantial variability across cities of similar size, which cannot be obtained from the deterministic model. The reason is that the process of disease transmission is inherently stochastic rather than deterministic. Hence, we now explore a stochastic version of the SIS model presented in the previous section. This will allow us to test whether diseases that transmit with less ease will also feature higher levels of variation at a given population size, as seen in the data.

In this scenario, we again assume that the contact rate c increases with population size as $c = bN^a$, the transmission rate is given by $\beta = pc$, where p is the per-contact infection risk and c is the number of contacts per unit of time. The probability a susceptible individual becomes infected between time t and $t + dt$ is given by

$$\mathbb{P}_I(t) = 1 - \exp[-\beta I(t)dt] = 1 - \exp[-pbN^a I(t)dt], \tag{13}$$

where $I(t)$ is the number of infected individuals in the population at time t, and N is the population size.

To recreate the variability seen in Figure 5, we ran simulations using the stochastic version of the SIS model presented above. Figure 6 shows the results of the SIS stochastic simulations. It shows a similar pattern to that of the empirical data in Figure 5 and the

deterministic SIS in Figure 2. Again, as the transmission rate increases, the scaling patterns become more linear, with a higher scaling exponent. Moreover, we also see in the figure that larger transmission rates are associated with lower variability, as seen in the empirical data. However, the variability in the simulations is considerably less.

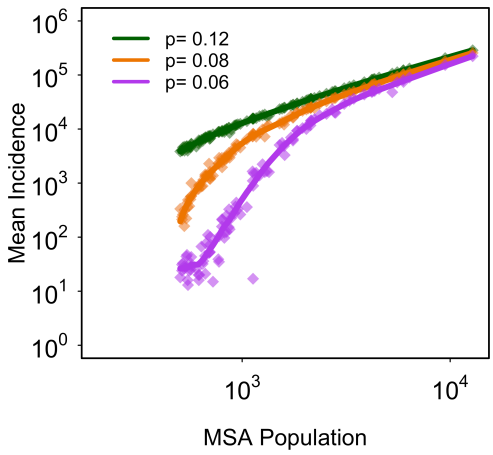

Figure 6: Incidence versus population, fitted with a loess line. Mean of ten stochastic simulations for each population size, with $\gamma = 0.3$ and $dt = 0.1$.

This stochastic model helps characterize, for a given phenomenon, the range of variation due to pure randomness. In other words, when one wonders whether a city has an unusually high incidence by pure bad luck, one wants to have an expectation for how (un)lucky can a city get. This stochastic version of the model suggests that the wide variation in the data is not simply pure randomness, but rather that the variation itself has a structural origin, which can be understood from the perspective of this SIS model.

5.4. The three aspects of complexity

A central mathematical relation of the theory derived in [23] states that:

$$E[Y_{c,f}] = Ne^{-M_f q(1-r_c)}, \tag{14}$$

where $E[Y_{c,f}]$ is the expected number of individuals engaged in a given phenomenon f (e.g., becoming infected with syphilis) in a city of population size N_c. The parameter M_f is the number of necessary elements a given urban activity (e.g., the transmission of an infectious disease) needs in order to occur. The parameter q is the probability any given individual

needs any of the M_f elements to be counted in said activity, and it is associated with the lack of susceptibility of individuals (e.g., to become infected). Lastly, r_c represents the diversity of elements city c, with size N_c, provides and, more precisely, it is the probability that an individual encounters any given element while living in the city. Thus, r_c quantifies the size of the social, economic, and cultural repertoire that is available in a city and to which individuals are exposed.

The link between Equation (14) and the power-law expression in Equation (1) emerges when r_c grows logarithmically with population size as $r_c(N_c) = a + b\ln(N_c)$. This logarithmic scaling of urban diversity, we posit, is a result that emerges from a selection and cumulative process of cultural evolution [26, 27]. Substituting this functional form in Equation (14) yields:

$$E[Y_{c,f}] = e^{-M_f q(1-a)} N_c^{M_f qb+1}. \tag{15}$$

Comparing expression (15) with the standard scaling model, $Y = Y_0 N^{\beta}$, the baseline prevalence Y_0 is given by $e^{-M_f q(1-a)}$, and the scaling exponent β by $M_f qb + 1$ (see [23] for more details). Note also the interesting similarities between expression (15) and the statistical model in (3).

The exponent in (14) has three terms that are proportionately related to the *complexity* of the phenomenon or activity, in the sense that increasing any of the terms decreases the probability that any given individual is engaged in the said phenomenon:

- 1) the *inherent* complexity of the phenomenon, which is represented by M_f
- 2) the lack of susceptibility of individuals to be engaged with the phenomenon, represented by q
- 3) the unsuitability of the system/environment for the phenomenon to occur, given by $1 - r_c$.