Long Range Dispersal Kernels; Theory and Applications

Chris B. Dock

***Abstract*—For the past year I have been exploring population physics as part of Professor Hallatschek's research group. My work has been for the most part contained within the field of spatially structured evolution, but has consisted in both theoretical explorations and simulations. My first task was to understand the language and mathematics of Population Physics, which I had never studied. To this end I relied heavily on [6] and [5]. I present an overview of the basic theoretical concepts of Population Physics in section I. My specific project was a continuation of [10] with two goals in mind: testing the models therein proposed on simulations and data from Google's Flu Trends API[18] (described in [8]), and studying certain theoretical results which would connect the problem to a generalized version of the celebrated Polya's Urn model.[1][11] With that in mind, I went through [10] in great detail and repeated its calculations, making sure to do out in full any result that I did not initially understand. This process is presented in section II. In section IV, I present an overview of the Polya's Urn problem followed by my Monte Carlo results. Finally, I present my attempts to apply the techniques of [10] to Google's Flu trends data in section V and conclude with an analysis of the possible usefulness of Google's Trends API for researchers studying dispersal dynamics.**

I. Introduction to Population Genetics

While I focused my study on the dynamics of dispersion, it should be understood in the context of spatially structured evolution. In this section I will build up some of the modern theory of evolutionary dynamics, following the introductory chapters of [9] and [6], and specify the precise extension to be studied in the remaining sections. Any understanding of evolutionary dynamics begins with the logistic equation,

$$\frac{df}{dt} = sf(1-f) \tag{1}$$

Where s is the fitness – the rate at which the frequency f would exponentially grow if it were not capped by the $1-f$ term. Note that (1) forces initial conditions within $[0,1]$ to remain so. As an interesting aside, (1) exhibits period doubling beyond $s=3$ and a transition to chaos for $s \approx 3.5699$. In evolution, however, s is the relative fitness of whatever allele is described by f and so is usually small enough to avoid this problem. The theory is thus far deterministic, which nature and reproductive success are not. The random chance component of reproductive success is termed genetic drift, and is typically incorporated by upgrading the frequency to a time dependent random variable and representing the drift as an uncorrelated noise term:

$$\frac{df}{dt} = sf(1-f) + \sqrt{\frac{f(1-f)}{N}}\eta(t) \tag{2}$$

By uncorrelated I mean η satisfies $\langle \eta(t) \rangle = 0 \ \forall t$ and that $\langle \eta(t)\eta(t') \rangle = \delta(t-t') \ \forall t,t'$. (2) is a type of stochastic differential equation known as a Langevin equation, familiar from statistical mechanics. Langevin equations are one of many equivalent ways of describing the dynamics contained in (2); another useful way is to view (2) as a hierarchy of coupled ODE's for the statistical moments of f. With this second interpretation in mind and following [9], it is conceptually important that for a neutral mutation ($s=0$) $\frac{d}{dt}\langle f \rangle = 0$. Hence, for neutral drift $\langle f \rangle = f$ and specifically $\lim_{t\to\infty} f = \lim_{t\to\infty} \langle f \rangle = p_{\text{fix}}$.[9] More generally, one could make (2) deterministic by considering instead the Fokker-Planck equation for the distribution of f as a function of time – in which case the noise coefficient becomes a frequency dependent diffusion-like coefficient.

From this simple starting point there are several possible directions of in which we might be interested in complicating (2). One direction, making f a vector of competing allele frequencies, leads to stochastic theories of ecological competition (as well as stochastic generalizations of the famous Lotka-Volterra predator-prey equation).[2] Another, making f a function of the base-pair vector $f(\vec{g},t)$ and adding mutation operators $\mu_{\vec{g}\to\vec{g'}}$ leads to theories of microscopic evolution or allele dynamics.[9] In fact this second direction is really a generalization of the first; the two are equivalent except that we don't generally observe animals to switch species. Finally, the direction of complication that I studied was the addition of spatial structure to the frequency: $f(\vec{x},t)$. There are several choices for the functional dependence on $\vec{x}$ of the right hand side of (2), which codify the spatial environment of the dynamics. We could for example, choose to make the relative fitness or the genetic drift magnitude a function of position, analogous to adding external forcing to a statistical mechanics problem. We will forgo external forcing, however, and instead add a term that allows the frequency at a given site $\vec{x}$ to influence the frequency of all other sites:

$$\frac{\partial f}{\partial t} = sf(1-f) + \int_{\mathbf{R}^d} G(|\vec{x}-\vec{y}|)f(\vec{y})d^d y + \sqrt{\frac{f(1-f)}{N}}\eta(t) \tag{3}$$

Where G, referred to as the jump-kernel, can be interpreted as the probability per unit space-time volume of a distance r jump occuring. Note that G is not confined to be a function but may be any tempered distribution (see [17]). Note that this model is sufficiently general to contain diffusion; consider

$$G(r) \equiv D[\delta''(r) + \frac{2}{|r|}\delta'(r)] \tag{4}$$

Then

$$\int_{\mathbf{R}^d} G(|\vec{x}-\vec{y}|)f(\vec{y})d^dy = D\int_{\mathbf{R}^d}(\partial_r^2 + \frac{2}{|r|}\partial_r)\delta(r)f(\vec{y})d^dy \tag{5}$$

$$= D\int_{\mathbf{R}^d} \nabla_y^2\delta(|\vec{x}-\vec{y}|)f(\vec{y})d^dy \tag{6}$$

$$= D\int_{\mathbf{R}^d} \delta(|\vec{x}-\vec{y}|)\nabla^2 f(\vec{y})d^dy \tag{7}$$

$$= D\nabla^2 f(\vec{x},t) \tag{8}$$

This somewhat odd jump-kernel reproduces diffusion exactly, but in fact any jump-kernel that declines exponentially or faster will produce diffusion like dynamics.[10] One caveat of this model is that it is not, for example, sufficiently general to contain the Eden model of spreading – in which terms of the form $|\nabla f|^2$ appear and make expansion normal to the boundary preferential.[13]

With the above motivations, a major component of the simulation work was to isolate the spreading dynamics of (2) and express

$$\frac{\partial f}{\partial t} = \int_{\mathbf{R}^d} G(|\vec{x}-\vec{y}|)f(\vec{y})d^dy \tag{9}$$

as a Monte Carlo simulation. Apart from the fact that this integro-differential equation is analytically difficult, there are theoretical motivations for again upgrading f to a random variable – this time based on the interpretation of G as the probability density of a given jump length. Specifically, jump-kernels with shallower tails, the dynamics is dominated by rare but extremely long range jumps.[10] This fact is not captured by (9) or indeed by (3), and so we must instead consider higher moments of the random variable generated by upgrading (9) to a stochastic process. Specifically, $G(r)$ is taken to be a random variable that takes the value $1/(dVdt)$ with probability $G(r)/\int G(r)dr$ and the value 0 otherwise (to jump or not to jump). The nature of the stochasticity therein produced is unfortunately somewhat less tractable than that in (2), mainly because the integral sign does not commute with the process of computing a statistical moment.[12]

As a final point concerning the overall theory, (3) is in fact a specific case of the multiple allele dynamics described in equation 1.51 of [9]:

$$\begin{aligned}\frac{\partial f(\vec{g})}{\partial t} =& \left[s(\vec{g}) - \sum_{\vec{g}'} s(\vec{g}')f(\vec{g}')\right] \\ &+ \sum_{\vec{g}'}\left[\mu_{\vec{g}'\to\vec{g}}f(\vec{g}') - \mu_{\vec{g}\to\vec{g}'}f(\vec{g})\right] \\ &+ \sum_{\vec{g}'}[\delta_{\vec{g},\vec{g}'} - f(\vec{g})]\sqrt{\frac{f(\vec{g}')}{N}}\eta(\vec{g}')\end{aligned} \tag{10}$$

Where $\vec{x},\vec{y} \to \vec{g},\vec{g}'$ and the integral discretizes into a sum. I mention this because it follows that the stochastic generalization of the jump-kernel described above might be a useful way to course grain the data contained in the mutation function $\mu_{\vec{g}\to\vec{g}'}$. The most profound difference between the two scenarios is not the discretization but the asymmetry that allele mutations often exhibit, in general $\mu_{\vec{g}\to\vec{g}'} \neq \mu_{\vec{g}'\to\vec{g}}$. One could account for this by letting

$$G(r)f(\vec{y}) \to G_{y\to x}(r)f(\vec{y}) - G_{x\to y}(r)f(\vec{x}) \tag{11}$$

II. Explanation of Central Problem and Explicit Calculations of [10]

In [10] the jump-kernels considered are power laws in $1/r$. It is shown that the dispersal exhibits two new qualitatively different behavior for different values of the power law exponent, and moreover that in intermediate time scales the dynamics are dominated by the marginal case between these two regimes. In this section I will describe the conceptual framework developed in [10] and carry out those derivations either present or implied therein that are necessary to produce the classification of power law jump-kernel behavior.

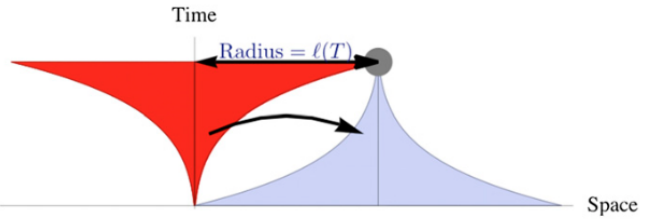

Figure 1. Taken directly from [10]. Graphical description of time-reversal symmetry argument.

The central argument concerns the consistency of the bulk 'growth funnel' generated by the dynamics in (9). The argument is laid out graphically in Figure 1: in order for the red funnel to be consistent, it must have a population of order one at its edge for any given time (e.g. the gray dot in Figure 1). It should be clear that in order for this to be the case, approximately one jump must have occurred from the red cone to the blue cone in the time before $t = T$. Written mathematically,

$$\int_0^T dt \int_{\mathcal{B}_{l(t)}}\int_{\mathcal{B}_{l(T-t)}} G(|\vec{y}-\vec{x}|)d^dxd^dy \sim 1 \tag{12}$$

Where the first ball $\mathcal{B}_{l(t)}$ is centered around zero (source funnel) and the second around an otherwise arbitrary point of distance $l(T)$ from zero (target funnel). Note that in this case G is to be interpreted as the probability per time per *source* volume per *target* volume of jumping a distance r. Assuming sufficiently fast growth, the difference $|\vec{x}-\vec{y}|$ will be approximately $l(T)$ (the separation between the cones dominating the placement of $\vec{x}$ and $\vec{y}$). Computing the spatial integrations we find

$$\int_0^T l^d(t)l^d(T-t)dt \sim 1/G(l(T)) \tag{13}$$

For sufficiently fast growth (or indeed sufficiently high dimension), the integrand will be dominated by the contribution at $t = T/2$, leading to the approximation

$$T\cdot l^{2d}(T/2) \sim 1/G(l(T)) \tag{14}$$

One can formalize this intuitive approximation using Laplace's method, but hte result is the same. If G is a known function then (14) provides a recurrence relation for $l(t)$. This is precisely what we will use to understand the dynamics for

$$G \equiv \frac{\epsilon}{r^{d+\mu}} \tag{15}$$

Where $\epsilon << 1$ and μ will turn out to be the key parameter. In this case, we find the recurrence relation

$$\epsilon t \cdot l^{2d}(t/2) = l^{d+\mu}(t) \tag{16}$$

Following [10], we clean this up slightly by defining the crossover scales at which the first jumps of order $l(t)$ will be made, i.e. when

$$\begin{aligned} \frac{tV_{\text{source}}V_{\text{target}}}{G} &\gg 1 \\ tl^{2d}(t)/G(l(t)) &\gg 1 \end{aligned} \tag{17}$$

We attain the crossover time and the crossover length separately by using the somewhat subtle fact that *before* the crossover scale, the dynamics are near-diffusive and adequately represented by $l(t) \approx v_0 t$. Note that by diffusive I do not mean $l(t) \sim t^{1/2}$ – an analysis due to the emminent statistical biologist Ronald Fisher shows that (3) with the jump-kernel given by (4) contains wave fronts with $v_0 = 2\sqrt{Ds}$. Using (17) and $l(t) \approx v_0 t$ one finds

$$t_\times = [\frac{v_0^{\mu-d}}{\epsilon}]^{1/(d+1-\mu)} \tag{18}$$

$$l_\times \equiv l(t_\times) = v_0 t_\times = [\frac{v_0}{\epsilon}]^{1/(d+1-\mu)} \tag{19}$$

For times much longer than the crossover scale, i.e. the asymptotic limit, we can safely represent (16) in units of the crossover scales. Specifically, defining $\lambda = l/l_\times$ and $\theta = t/t_\times$ gives

$$\lambda^{2d+\delta}(\theta) \sim \theta\lambda(\theta/2)^{2d} \tag{20}$$

In which, anticipating the critical value of μ from the crossover scales, we have defined $\delta = \mu - d$. This is is a recurrence relation not in step but in scale, suggesting the use of logarithmic variables $\phi = \log_2 \lambda$ and $z = \log_2 \theta$ so that $\theta/2 = 2^z/2 = 2^{z-1} = \theta(z-1)$. Computing the logarithm of (16) gives

$$(2d+\delta)\phi(\theta(z)) = z + 2d\phi(\theta(z-1)) \tag{21}$$

Which is a non-homogeneous difference equation for $\phi(z) \equiv (\phi \circ \theta)(z)$. In order to eliminate the non-homogeneous term z it suffices to include a linear particular solution:

$$\phi(z) = \phi_{\text{hom}}(z) + \alpha z + \beta \tag{22}$$

Inserting this ansatz into (21) gives

$$(2d+\delta)\phi_{\text{hom}} = 2d\phi_{\text{hom}} + (-\delta\alpha + 1)z + (-\delta\beta - 2d\alpha) \tag{23}$$

From which one obtains

$$\begin{aligned} (2d+\delta)\phi_{\text{hom}} &= 2d\phi_{\text{hom}} \\ \alpha &= \frac{1}{\delta} \\ \beta &= -\frac{2d}{\delta^2} \end{aligned} \tag{24}$$

The homogeneous equation for $\phi_{\text{hom}}(z)$ can be solved via the standard ansatz $\phi_{\text{hom}}(z) = \phi_{\text{hom}}(0)\Lambda^z$, yielding

$$\phi_{\text{hom}}(z) = \phi_{\text{hom}}(0)(1 + \frac{\delta}{2d})^{-z} \tag{25}$$

And finally

$$\frac{\delta^2}{2d}\phi(z) \approx \frac{\delta z}{2d} + (1 + \delta/2d)^{-z} - 1 \tag{26}$$

This is (S9) from [10], and it contains the asymptotic behavior for $l(t)$. As we shall see, the asymptotic case is suppressed for intermediates times by the critical $\mu = d$ case. (26) can be simplified if the value of μ is known a priori, and doing so reveals how the qualitative behavior depends on μ. The first and second terms dominate respectively for $\delta > 0$ and $\delta < 0$ as $z \to \infty$, giving:

$$\log(\lambda(t)) \approx \begin{cases} B_\mu t^\eta,\ \eta = \log \frac{2d/(d+\mu)}{\log 2} & \delta < 0 \\ \log(A_\mu t^\beta),\ \beta = \frac{1}{\mu - d} & \delta > 0 \end{cases} \tag{27}$$

With the prefactors obtainable by allowing (26) to pass to the appropriate limits:

$$\begin{aligned} \log A_\mu &= -2d\log(2)\delta^{-2} \\ B_\mu &= 2d\log(2)\delta^{-2} \end{aligned} \tag{28}$$

Note that (27) also contains the linear and sub-linear regimes that for $\mu > d+1$. Thus, we have shown that the core radius exhibits diffusive behavior for $\mu > d+1$, power-law behavior for $d+1 > \mu > d$, and stretched exponential behavior for $d > \mu$.

Finally, it is a critical point that thus far the case $\mu = d$ has hitherto been ignored. It is clear from (28) that there will be convergence problems as $\delta \to 0$ since both prefactors contain δ^{-2}. (24) reveals the source of the problem – the linear particular solution to (21) does not work when $\delta = 0$. We seek instead a quadratic solution of the form $\phi(z) = az^2 + bz$. No constant term is necessary because the coefficients of $\phi(z)$ and $\phi(z-1)$ are equal, canceling any constant term – this arbitrary constant will reappear in the full solution as $\phi(0)$. With $\delta = 0$, (21) may be written

$$\phi(z) = \phi(z-1) + \frac{z}{2d} \tag{29}$$

Which, plugging in our new ansatz, gives

$$\begin{aligned} az^2 + bz &= a(z^2 - 2z + 1) + b(z-1) + \frac{z}{2d} \\ a = b &= \frac{1}{4d} \end{aligned} \tag{30}$$

A glance at (24) in the case $\delta = 0$ reveals that the homogeneous solution is simply $\phi(0)$. Thus our new full solution is

$$\phi(z) = \phi(0) + z(z+1)/4d \approx \frac{z^2}{4d} \tag{31}$$

Thus the necessary addendum to (27) is

$$\log(\lambda(t)) \approx \begin{cases} B_\mu t^\eta,\ \eta = \log \frac{2d/(d+\mu)}{\log 2} & \delta < 0 \\ \frac{\log^2(t)}{4d\log(2)} & \delta = 0 \\ \log(A_\mu t^\beta),\ \beta = \frac{1}{\mu - d} & \delta > 0 \end{cases} \tag{32}$$

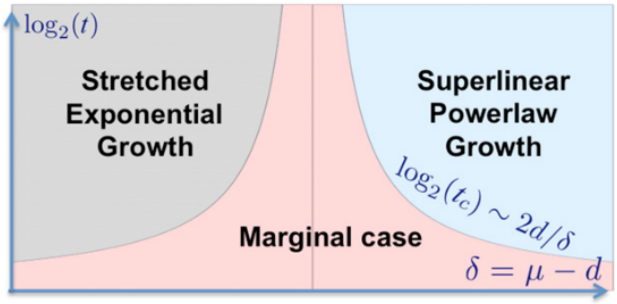

Figure 2. Taken directly from [10]. Demonstrates the interference of the critical case in intermediate time scales.

As is shown in [10], the spreading of the core radius is essentially given by the critical $\delta = 0$ unless

$$\log_2(t) \gg \frac{2d}{|\delta|} \text{ and } \log_2(l) \gg \frac{2d}{\delta^2} \tag{33}$$

This is demonstrated pictorially in Figure 2.

I took a somewhat sideways approach to studying the transition between the marginal case and asymptopia: By considering a threshold of de-correlation between $\lambda_{\text{critical}}(t)$ and the observed radial growth one can evaluate which values of μ cease to feel the influence of the critical case. By Figure 2 this is equivalent to finding the times beyond which the marginal case is irrelevant for a given μ (one simply has to reflect about the symmetry axis of the separating hyperbola). Plotted in Figure 3 is the R^2 value of fitting the theoretical critical curve to the data via scaling. I expected the resulting curve to be nearly symmetryic about $\mu = d$, and was shocked that it was not.

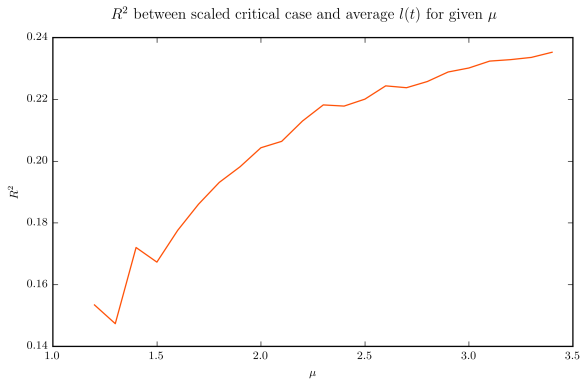

Figure 3. R^2 value of scaling fit fit of $l_\mu(t)$ vs. μ. I was shocked by this plot. It seems the picture is much less symmetric than was shown in Figure 2, and that the critical case extends its influence far into the powerlaw and linear regimes.

With the three regimes of (32) in mind, I set about testing the results derived above on Monte Carlo simulations, as well as using (13) to see if I could extract μ from noisy $l(t)$ data.

III. Simulations Testing Funnel Model of [10]

Once I had a decent theoretical understanding of the constructions in [10] I set about building a platform to run and parallelize simulations, as well as to automate the plotting of the outputs. I wrote management scripts in Python to initiate simulations (sometimes in parallel) and analyze their outputs, but the Monte Carlo simulations themselves were written in C. C has the enormous speed advantage that accompanies compiled languages[14], and since I was often working on my own computer the factor 40 speedup was essential. The trade off of course is that C has very few built in functions and data structures; before I could begin I had to implement hash tables, dynamic arrays, and a random number generator. I used the Mersenne Twister algorithm to produce random numbers, which has been shown to be more than adequate for non encryption-grade work.[15] My full source code and documentation are available at [3].

I chose to work in dimension $d = 2$ both because it is the relevant dimension for understanding epidemics and because it is the lowest dimension in which non-trivial sector formation is possible. More generally, 2 is the lowest dimension for which the configuration geometry can vary significantly, making the problem of multiple alleles much more interesting. The dynamics and stability of sectors in spreading phenomena has recently been of interest in understanding the frequency of positive mutation events in expanding populations, e.g. drug resistance in bacterial colonies.[4] I examine sectors in the long-range dispersal regime below.

The end results of three typical simulations are shown in Figure 4. One thing to note is the colors of these images indicate age in 'generations', which scales as $\log_2(\text{population})$. I use the word generation because if one were to count up the number of sites with g ancestors or fewer one would get approximately 2^g. The fact that the rightmost image is dominated by yellow means that nearly all of the population is *recent* – which will of course always be the case in the stretched exponential growth regime. One can tell fairly clearly from these plots that the low μ regimes have dynamics dominated by rare (but not too rare) jumps of order $l(t)$. Figure 5 illustrates this further. The plots in 5 also contain the best fits of the critical regime theoretical curve to the average radius curves. I say 'fit' because there is an unknown parameter in all three of our studied regimes, namely an overall factor due to the crossover scale.

The accuracy of the critical theory in different regimes is of some interest, as it indicates the relative magnitudes of the correction terms left out of (32) as they vary with δ. Moreover, the critical regime was shown earlier to dominate asymptopia for significant time scales, and hence we expect that the deviation will grow as the asymptotics "switch on." For example, it can be seen in Figure 6 that the critical approximation does extremely well for $\mu = 2.5$, and *less* well for $\mu = 2.0$ and $\mu = 1.5$. This is surprising since $\mu = 2.0$ is in fact the critical case, and it indicates that in intermediate time scales there are significant corrective terms to either the powerlaw regime, the critical regime, or both. In Figure 3 I plot the R^2 value of the critical regime fit against μ, demonstrating the asymmetry in δ of the corrections to the first order theory classified in Figure 2.

I then set out to use (13) to estimate μ from the radial growth data of Monte Carlo simulations. By convolving $l^d(t)$ with itself, one obtains a curve for $(G \circ l)(t)$ which can

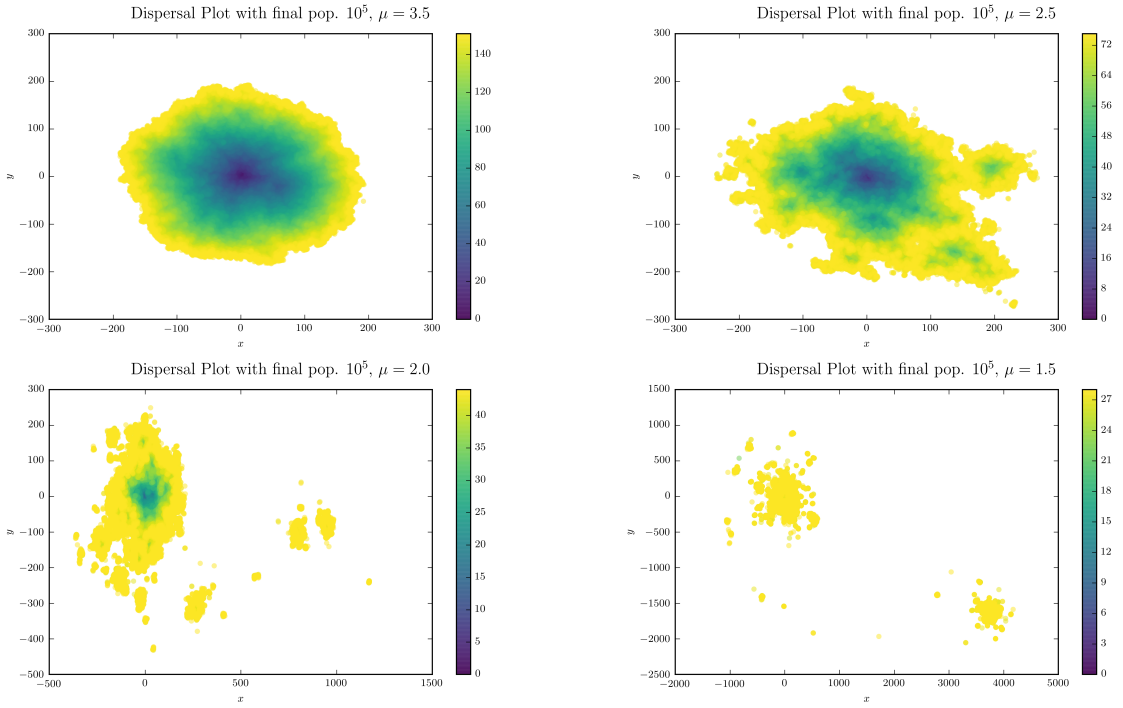

Figure 4. The spatial outputs of Monte Carlo simulations. In reading order, these simulations had $\mu = 3.5$, $\mu = 2.5$, $\mu = 2.0$, and $\mu = 1.5$. Final population in each simulation was 10^5 lattice sites. The color indicates age, blue meaning old and yellow meaning new.

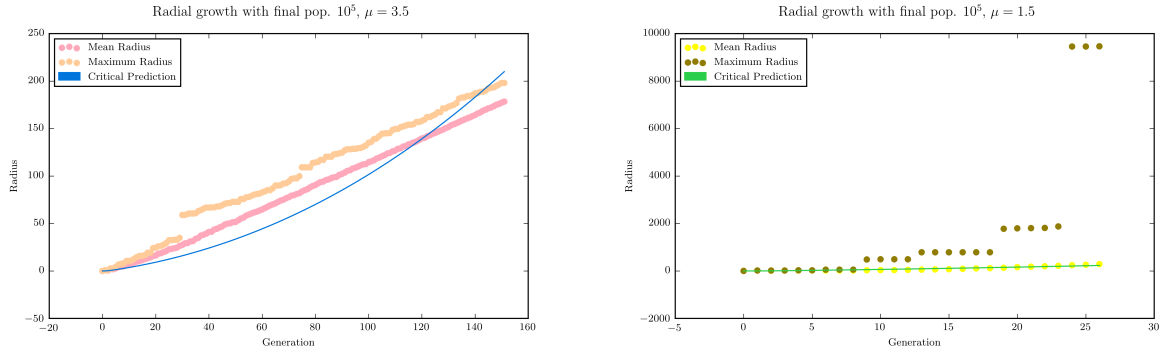

Figure 5. Shown here are two different measures of the radius as a function of time – the distance from the origin to the furthest occupied site and the average distance in a given generational unit. On the left is the linear growth regime and the right the stretched exponential regime. Clearly, 'rare' jumps are mostly unimportant in the diffusive case and diffusive motion is mostly irrelevant in the stretched exponential case.

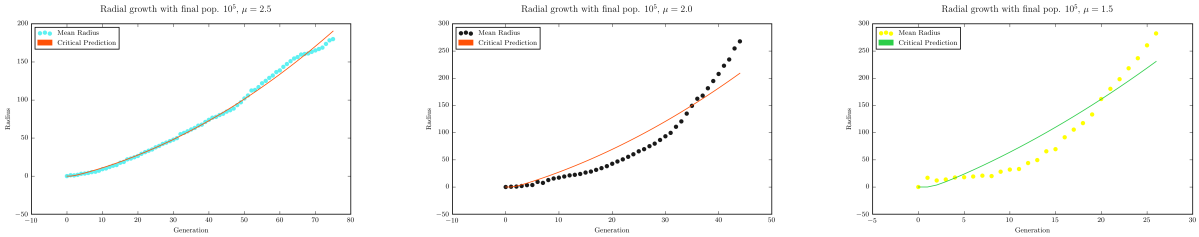

Figure 6. The radial growth of three Monte Carlo simulations with μ decreasing from left to right, is compared with the theoretical expectation given by the critical regime (which is expected to dominate the asymptotic regime initially).

be fit for μ using our power law assumption for the jump-kernel. Confirming that this was possible for simulations with known μ was necessary before I could attempt using the data from Flu Trends. My results were mixed – in two dimensions several hundred large simulations were necessary before the μ values would appear to converge. The extra dimension vastly increases the number of possible motions for the spreading, and as such the variability in any estimate of μ. An example of the estimates therein produced, and of their inadequacies, is shown in Figure 7. Clearly, some signal of μ is present in the Monte Carlo outputs but it appears to be systematically shifted downwards. I should note that in the case of Figure 7 and in most instances that I have observed, the process of extracting the μ values from data preserves their ordering. It's unclear to me why there is this systematic shift, but since it is so regular it could easily be accounted for in studying real world data.

Another I technique I used was to exploit the fact that the dynamics generated by (9) are inherently recursive; the initial growth funnel spawns seedlings, which then evolve independently from the spawning funnel for a characteristic amount of time before the two collide. By capturing these sub-funnels, I was able to increase the number of $l(t)$ data-points available to the μ extraction effort by about 10% (See Figure 8).

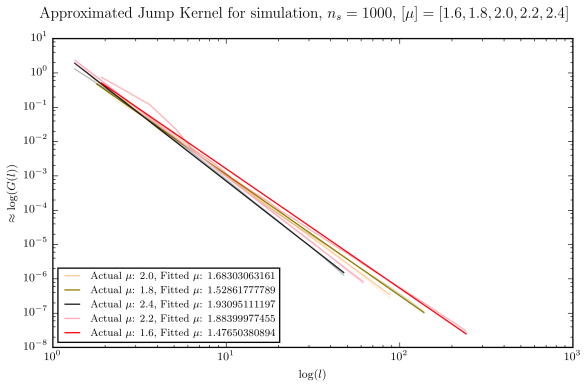

Figure 7. Attempt to use (14) to estimate μ from Monte Carlo output. Data from $n_{\text{sim}} = 1000$ simulations of final population 10^5 is averaged for each μ, but the results are lower than expected – ie I am observing faster than expected growth, increasingly so as μ increases.

The addition of secondary funnels helped, but the μ values I was estimating still varied wildly and seemed down-shifted on average. Predicting the variation in estimated values of μ is difficult because the technique used to produce the estimates is to perform a numerical version of the convolution in (14) on noisy data and then to fit an inverse power law to the resulting kernel. Moreover we are fitting for a derivative in log-log space, so least squares performs minimization performs poorly. I chose to use a log likelihood maximization library provided by [7]. Since I was unsure how to examine the variability of μ_{approx} analytically I resorted to numerical experiments. See Figure 9 for a histogram of μ values produced for a given set of simulation parameters.

I next returned to the question of configuration geometry that I mentioned earlier as a reason for choosing to work in 2

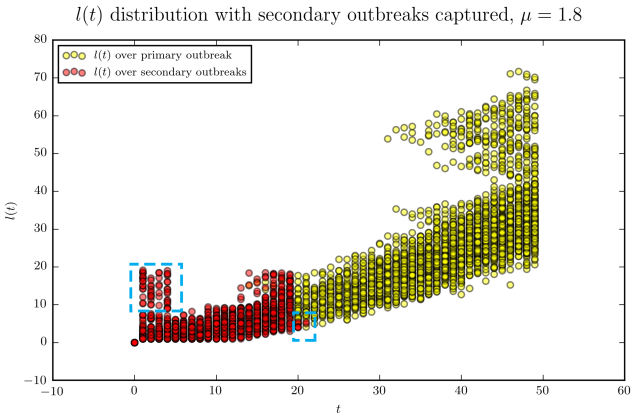

Figure 8. The time axis is in units of generation, above each value of g are all of the different distances from the origin of sites in that generation. Each red point (overlayed) is taken from a secondary funnel and measures their radius and generation relative to *it* and not the central funnel. I have highlighted certain problem areas in the distribution of red points that result from collision with other funnels.

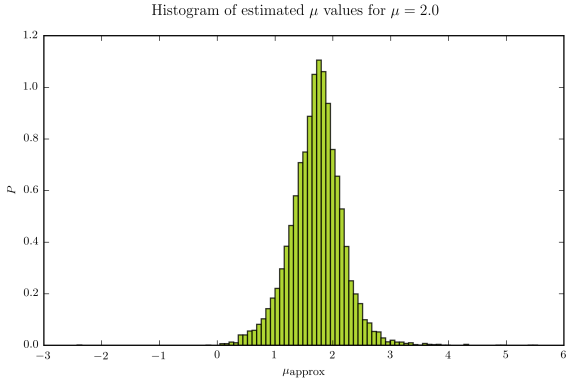

Figure 9. A histogram of estimated μ values using the convolutional technique described above. In this case the correct value of μ was 2.0, and we see again the fact that the method is undershooting the correct value. Moreover, the distribution has fairly long tails, making predictions based on single samples of real world data difficult.

dimensions. By configuration geometry I simply mean the initial conditions of the Monte Carlo simulation or equivalently of (9) – in all of the aforementioned simulations the configuration was chosen to be a single point. For a single allele modifying the configuration geometry doesn't do anything interesting; one only has to wait a short amount of time for the dynamics to start acting like they would starting from a point. For two or more alleles however, geometry becomes extremely important. The different mutants can become trapped, or they can escape, or they can interfere with each other – the probabilities of these events being sensitive to the initial configuration.

The two fundamentally distinct initial topologies that I considered are the bubble and the sector (noting the importance of these two types to the effect of spatial structure on jackpot events as shown in [4]). I was interested in studying the relative dissipation times of these two types of geometries in the low μ regime. By defining a point by point Hamming distance and continually rescaling the funnel to have the same approximate area as its initial configuration I was able to define a 'mixing' time – a time after which the geometry has

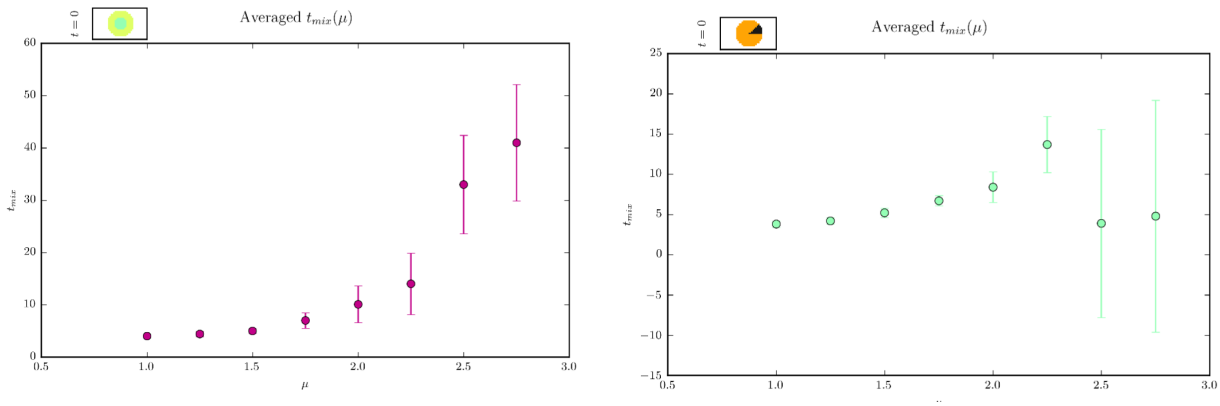

Figure 10. Mixing times for bubbles vs. sectors as a function of μ. No that sectors tend to curve outwards for high μ, destroying their initial geometry according to the Hamming metric but perhaps not according to some more useful choice of distance.

completely disappeared. The mixing times for bubbles and a specific sector angle are shown in Figure 11. An example of the kind of chance event which creates geometric dissipation is shown in

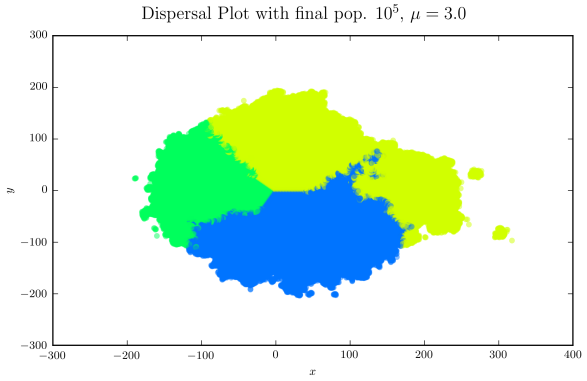

Figure 11. Geometric dissipation of an evenly split three sector structure. All that is needed is for an allele to jump into the sector of another allele, far enough out that it 'wins' their inevitable collision.

IV. Polya's Urn

Polya's urn is a classical statistics problem, essentially the opposite situation to sampling without replacement: Given an urn containing black and white (in fact, arbitrarily many colors) balls in some frequencies, a ball is taken out at random and $textsltwo$ balls of that color are put back. The typical question is then the variability of the relative frequencies in the limit $t \to \infty$. This problem is extremely well studied and many results have been obtained.[1][11] The connection to the problem at hand is essentially the limit $\mu \to 0$. As the expected jump length diverges (as it does for $\mu \to 0$), spatial extent becomes irrelevant to the dynamics, and we get the take one out put two back Polya's Urn process. One of the questions I studied, therefore, was the equivalent question regarding the variance in the quantities

$$F_i \equiv \lim_{t\to\infty} f_i(t) \tag{34}$$

Where f_i are the random variables representing the prevalence of the ith allele. Several examples are shown in Figure 12 for how the frequencies evolve in the multiple allele case – examples that I hope illustrate that the distribution of the F_i's is hard to predict by studying the short to medium term behavior. In order to get lower limit for how the variance of F_i depends on μ, I considered the variance of the quantity $F = \max_i(f_i)$. This variance is of course taken 'over histories', so I ran 10^5 simulations and computed the variance and mean of F at each generation (Figure 13). I needed a single quantity because, between histories, it is impossible to say which allele corresponds to which when they are identical. The mean of F is somewhere between $1/2$ and 1, as expected, and will steadily climb towards 1. Since the mean is climbing towards the maximum possible value, the distribution is being squeezed and the variance should decrease. We don't see this beginning to occur until 170 generations in, however. This curve should in theory provide a lower bound on the variance for the f_i's as a function of time.

V. Google Flu Trends

The final component of my project was to apply the techniques above to Google's Flu Trends data. I first wrote a program to, from a collection of search terms that I guessed to be correlated with flu incidence ('flu', 'fever', etc.), produced weights which most highly correlated the integrated Flu Trends data with the integrated data from the CDC (a more trustworthy but less high resolution source). The results are shown in Figure 14. Note that while the two are highly correlated, the Flu trends data (predictably) exhibits a 'virality' phenomenon. There was, for example, a tremendous panic over H1N1 before very many cases had occurred.

After matching the Google data with the CDC data on a national scale, I then turned to the spatially resolved picture which was available from search trends. Unfortunately, Google breaks the country up according to Nielson DMA's (advertising districts) – making for a fairly uneven density of data points (Figure 15).

Zooming in on a single ball to the one highlighted in Figure 15, this time around Baltimore, and restricting our time period

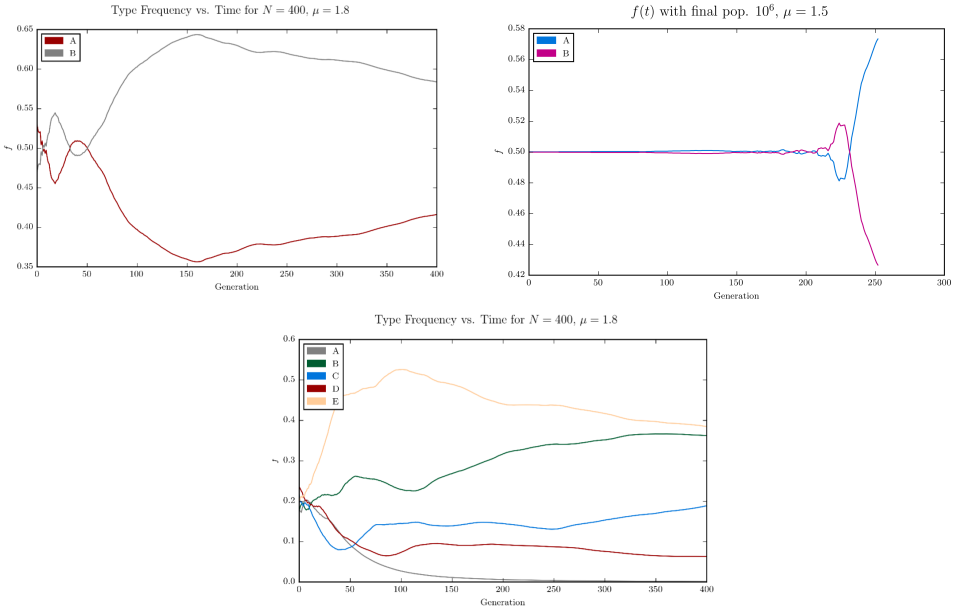

Figure 12. Several different scenarios for the evolution of allele frequencies. For simplicity, all start from a well mixed initial configuration. In the first two there are only two alleles present, and even so it is unclear what will occur even after 400 generations. In the third we have five different alleles, one of which gets unlucky and dies out almost immediately, but the distribution of the rest is unclear

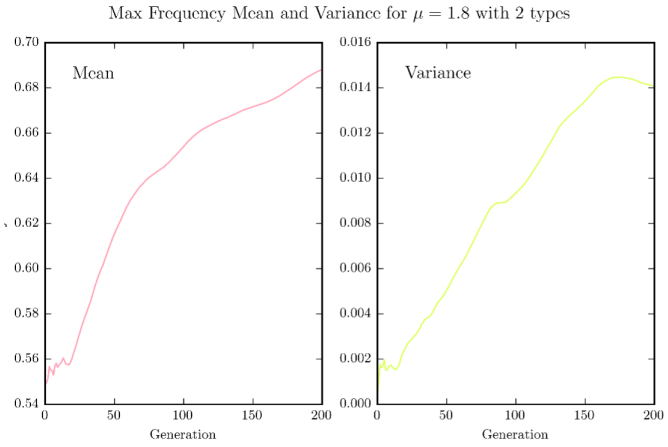

Figure 13. Mean and variance of the random variable represented by the largest frequency at any given time. Provides a lower bound for the variance of f_i, but not useful for infinite times because it (theoretically) goes to zero.

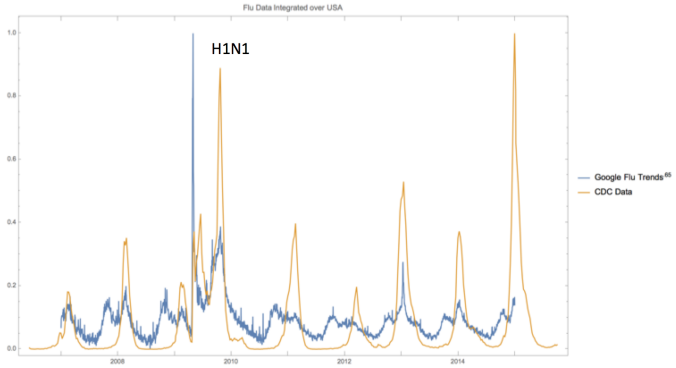

Figure 14. Rescaled CDC data and Google Flu Trends data with optimized search terms.

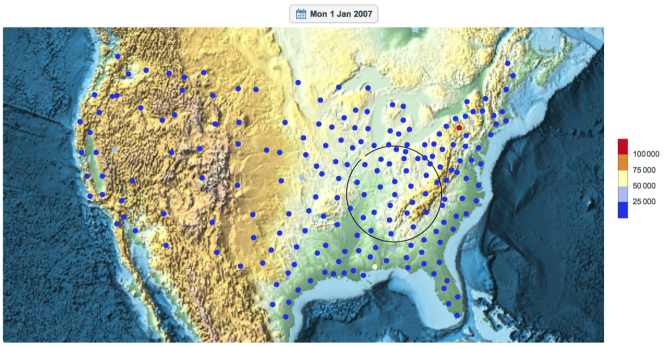

Figure 15. Map of the data points available via Google's Flu Trends, with a 400 mile radius ball around Greenville Kentucky pointed out as a fairly uniform density subset.

to a single flu outbreak in 2008, we find the integrated signal to be as in Figure 16. I restrict my analysis to balls of a certain radius based on the expectation that the de-correlation length will be small. We are not in interested in the integrated signal, however, but in its spatial structure. Thus we need to compare the 'times' that the signal approximately reached each DMA in the ball. Some of these DMAS have poor data, however (sometimes Google's data has large blank areas). I filtered these out using Mathematica's Classify capability. The Classifier learned what bad data looked like and filtered them out. The spatially dependent signals of the rest are shown in 17. From this point I needed to determine an approximate 'time' between the signals. In order to do this I minimized the function

$$d(\tau) = \int_{\text{Period}} (S_1(t) - S_2(t+\tau))^2 dt \tag{35}$$

This gave approximate time separations between τ_{ij} any two

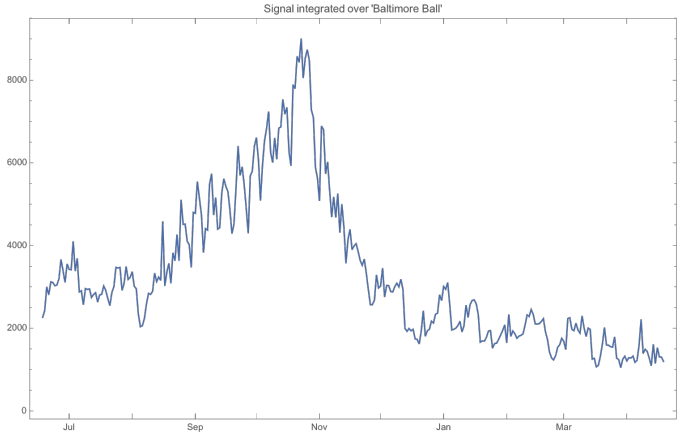

Figure 16. Signal integrated over ball around Baltimore, containing 10 distinct DMAS post filtering.

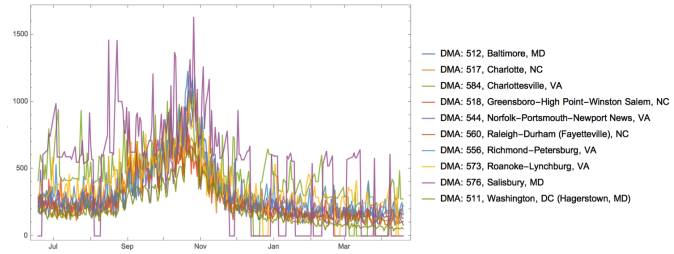

Figure 17. Spatially dependent flu signal, shown for each of the 10 'good' DMAs in the chosen region around baltimore.

of the DMA's with respect to the flu signal. I then converted this into a global picture for time by diagaonalizing τ_{ij}. This gave me a similar result to what I initially obtained by just computing the weighted mean time of each signal. With the time separations thusly constructed, it was possible to look at the space time picture (shown pictorially in Figure 1). Plotting the space time 'funnel' (with respect to a spatial coordinates of latitude and longitude, e.g. a Mercator projection) gave Figure 18. It's difficult to say whether any funnel like behavior is

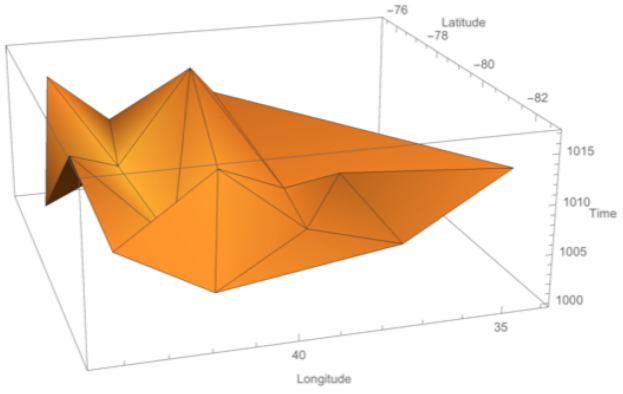

Figure 18. Spacetime plot of data in baltimore ball based on flu signal. Unclear whether any funnel like behavior is occurring.

occurring in Figure 18 – primarily because there are so few points.

Another factor which made me uncertain was that there

loops in the time separation matrix τ_{ij}, even for signals which were far apart from each other.

I repeated the above process for several different years and regions, and found similarly challenging results.I wanted to figure out whether this was a universal property of the flu spread, so I returned to the CDC data and applied (35) to discern τ_{ij} for the 9 CDC HHS regions (shown in Figure 20). The separate signals for each HHS region are shown in Figure 19, and the resulting τ_{ij} matrix in Figure 21. The key fact

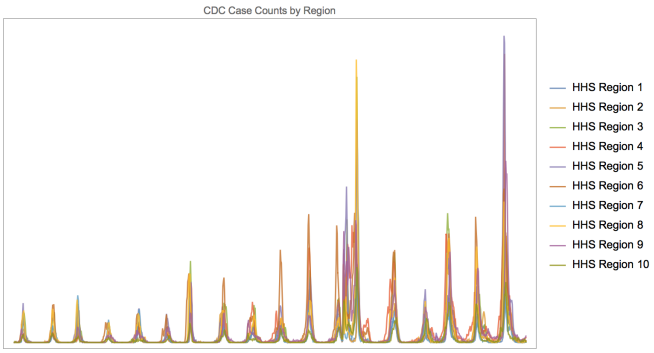

Figure 19. Data from the CDC by HHS region (Figure 21).

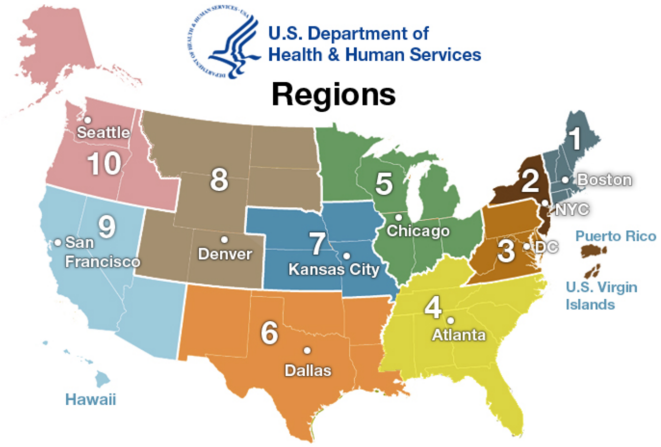

Figure 20. HHS Regions.

X	HHS 1	HHS 2	HHS 3	HHS 4	HHS 5	HHS 6	HHS 7	HHS 8	HHS 9	HHS 10
HHS 1	0.	0.058	1.358	3.647	2.386	3.42	2.649	2.741	0.509	1.962
HHS 2	−0.058	0.	0.806	2.549	1.92	2.774	2.195	2.327	−0.431	1.549
HHS 3	−1.358	−0.806	0.	1.23	0.737	2.	1.	0.977	−1.138	−0.136
HHS 4	−3.646	−2.549	−1.23	0.	0.055	0.406	−0.192	−0.917	−3.495	−2.248
HHS 5	−2.386	−1.92	−0.737	−0.055	0.	0.949	0.292	0.364	−3.691	−0.883
HHS 6	−3.42	−2.774	−2.	−0.405	−0.949	0.	−0.712	−1.04	−3.	−1.714
HHS 7	−2.649	−2.195	−1.	0.193	−0.291	0.712	0.	0.007	−1.717	−1.067
HHS 8	−2.741	−2.327	−0.977	0.917	−0.364	1.04	−0.007	0.	−1.809	−1.014
HHS 9	−0.507	0.433	1.139	3.495	3.691	3.	1.72	1.809	0.	0.542
HHS 10	−1.961	−1.548	0.136	2.247	0.883	1.714	1.067	1.014	−0.54	0.

Figure 21. τ_{ij} produced from CDC data on the national scale. Note that this matrix *does* contain loops: Region 7 was hit before .292 weeks before Region 5 which was hit .055 weeks before Region 4 which was hit .193 weeks before Region 7. Fortunately these regions are adjacent; as far as I can tell τ_{ij} does not contain the kind of extended loops that troubled me about the Flu Trends data near Baltimore.

concerning Figure 21 is that it does not contain extended loops,

and therefore does appear to be a spatial structure in the way the flu moves across the country. The fact that extended loops do occur in the Flu Trends data is troubling; it indicates either that the search data is too noisy or that there is something fundamentally different from flu signals about the way 'search signals' are transmitted.

VI. Conclusions

I would very much like to thank Professor Hallatschek and the entire research group for affording me the opportunity to explore Population Physics, a field that I had never previously encountered in my coursework. While I'm not convinced that Google's public Flu Trends API is high enough resolution to thoroughly test the models of disease spread specified in [10], I am nevertheless convinced of its usefulness to researchers studying general spreading phenomena because of the breadth of spatially correlated search terms available. I anticipate that further research will reveal that it contains phenomena spanning the full range of μ values, from linear waves to scale free and even uncorrelated examples. Higher resolution data would allow for the exploration of all the aforementioned phenomena in 'bumpy' terrain, with varying levels of resistance to spreading for a given search term. The increasing availability of this kind of data from Google and the CDC and others[16] makes it an exciting time indeed to be studying both evolution and spreading phenomena more generally.

REFERENCES

[1] Fan Chung, Shirin Handjani, and Doug Jungreis. Generalizations of polya's urn problem. *Annals of combinatorics*, 7(2):141–153, 2003.

[2] Ulf Dieckmann and Richard Law. The dynamical theory of coevolution: a derivation from stochastic ecological processes. *Journal of mathematical biology*, 34(5-6): 579–612, 1996.

[3] Chris Dock. Flu trends. https://github.com/cbartondock/flu_trends, 2017.

[4] Diana Fusco, Matti Gralka, Jona Kayser, Alex Anderson, and Oskar Hallatschek. Excess of mutational jackpot events in expanding populations revealed by spatial luria-delbruck experiments. *Nature communications*, 7, 2016.

[5] JL Garcia-Palacios. Introduction to the theory of stochastic processes and brownian motion problems. *arXiv preprint cond-mat/0701242*, 2007.

[6] Lukas Geyrhofer. *Quantifying evolutionary dynamics*. PhD thesis, Georg–August–Universität Göttingen, 2014.

[7] Adam Ginsberg. Powerlaw fit. https://github.com/keflavich/plfit, 2017.

[8] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[9] Benjamin H. Good. *Molecular evolution in rapidly evolving populations*. PhD thesis, Harvard University, 2016.

[10] Oskar Hallatschek and Daniel S Fisher. Acceleration of evolutionary spread by long-range dispersal. *Proceedings of the National Academy of Sciences*, 111(46):E4911–E4919, 2014.

[11] Fred M Hoppe. Pólya-like urns and the ewens' sampling formula. *Journal of Mathematical Biology*, 20(1):91–94, 1984.

[12] John Hunter. Stochastic processes.

[13] Pawel Keblinski, Amos Maritan, Flavio Toigo, Russell Messier, and Jayanth R Banavar. Continuum model for the growth of interfaces. *Physical Review E*, 53(1):759, 1996.

[14] Miles Lubin and Iain Dunning. Computing in operations research using julia. *INFORMS Journal on Computing*, 27(2):238–248, 2015.

[15] Makoto Matsumoto and Takuji Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30, 1998.

[16] Richard A Neher and Trevor Bedford. Nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, page btv381, 2015.

[17] Robert D Richtmyer and Wolf Beiglböck. *Principles of advanced mathematical physics*, volume 2. Springer, 1981.

[18] Google ™. Google public data.