# Modeling stochastic processes in disease spread across a heterogeneous social system

**Minkyoung Kim[a,1], Dean Paini[b], and Raja Jurdak[a,c,d]**

[a]Data61, Commonwealth Scientific and Industrial Research Organisation, Pullenvale, QLD 4069, Australia; [b]Health & Biosecurity, Commonwealth Scientific and Industrial Research Organisation, Canberra, ACT 2601, Australia; [c]School of Information Technology and Electrical Engineering, University of Queensland, St Lucia, QLD 4072, Australia; and [d]School of Computer Science and Engineering, University of New South Wales, Kensington, NSW 2052, Australia

Edited by Burton H. Singer, University of Florida, Gainesville, FL, and approved November 21, 2018 (received for review January 25, 2018)

**Diffusion processes are governed by external triggers and internal dynamics in complex systems. Timely and cost-effective control of infectious disease spread critically relies on uncovering underlying diffusion mechanisms, which is challenging due to invisible infection pathways and time-evolving intensity of infection cases. Here, we propose a new diffusion framework for stochastic processes, which models disease spread across metapopulations by incorporating human mobility as topological pathways in a heterogeneous social system. We apply Bayesian inference with the stochastic Expectation–Maximization algorithm to quantify underlying diffusion dynamics in terms of exogeneity and endogeneity and estimate cross-regional infection flow based on Granger causality. The effectiveness of our proposed model is shown by using comprehensive simulation procedures (robustness tests with noisy data considering missing or delayed human case reporting in real situations) and by applying the model to real data from 15-y dengue outbreaks in Australia.**

disease spread | Hawkes process | infection flow | human mobility

Diffusion processes in the real world often produce non-Poisson distributed event sequences, where interevent times are highly clustered in the short term but separated by long-term inactivity (1). Examples are observed in both human and natural activities such as resharing microblogs in online social media (2, 3), citing scholarly publications (4, 5), a high incidence of crime along hotspots (6, 7), and aftershock sequences near the seismic center (8). These all imply that an event occurrence is likely triggered by preceding events in cascades of different scales, and the timing of discontinuous events conveys information of underlying diffusion mechanisms.

Based on point process approaches, uncovering such feedback mechanisms between preceding and triggered events has drawn significant attention from a wide range of scientific communities (2–9), since it helps predict diffusion trends and establish cost-effective strategies for the promotion or restriction of the diffusion process (10). When it comes to epidemics, an accurate understanding of underlying dynamics is crucial for the timely control of infectious disease spread. However, uncovering disease dynamics is very challenging due to unobservable transmission routes and limited information of private social networks, unlike explicit cited–citing relationships of documents in online social media (2, 3, 9, 11) or in academic publications (4, 5). Moreover, large international and domestic travel volumes have increased the uncertainty of infection pathways. Thus, the quantification of exogenous and endogenous effects is essential to overcome the challenges and understand emergent bursts of outbreaks, and has been largely neglected in epidemic studies (12–14).

In this study, we propose the Latent Influence Point Process model (LIPP) for disease spread across a heterogeneous social system by incorporating three major counterbalancing factors: (*i*) exogenous influence covering environmental heterogeneity, (*ii*) endogenous influence attributed to macrolevel interactions between metapopulations, and (*iii*) a time decay effect. We apply Bayesian inference using the stochastic Expectation–Maximization algorithm, which enables us to quantify the reflexivity of metapopulations, i.e., the level of feedback on event occurrences (15, 16), driven by external and internal dynamics in a complex system, and to estimate infection flow between metapopulations based on Granger causality.

We first conduct simulations to generate synthetic data as ground truth by varying parameter settings so as to mimic real data. We also add variations to the generated datasets for reproducing (*i*) random missing, (*ii*) nonrandom missing, and (*iii*) time-delayed mechanisms of human case reporting. With 1,200 datasets in total, we evaluate the model performance (recovery of infection flow between regions infection flow and model parameters) and compare with competing baselines. Our model well recovers cross-regional infection flow, with greater than 85% accuracy (over 70% for noisy data) with a 95% confidence interval, and outperforms baseline models. For real data, we investigate dengue spread in Queensland, Australia, during a 15-y period (2002–2016). We find that dengue outbreaks become more globally interconnected across multiple regions through human mobility, leading to more complex behavior of disease spread over time. In terms of reflexivity, precursory growth and symmetric decline of outbreaks in metropolitan or populated regions are attributed to slow but persistent feedback on

## Significance

**This study infers probabilistic infection routes of a vector-borne disease, by modeling internal dynamics of metapopulations driven by human mobility as multivariate stochastic processes. In this way, our proposed model uncovers the self-excitation and mutual excitation nature of disease spread across a heterogeneous social system with rich context. Our model is a general extension of networked Hawkes processes, providing flexibilities to add constraints (presence of diffusion medium) and to use domain knowledge (cross-metapopulation connectivity), enabling covering of direct and indirect diffusion processes such as contact-based and vector-borne disease spread. Our model is readily applicable to a wide range of intragroup and intergroup diffusion processes in social and natural systems and can infer probabilistic causality between discrete events.**

Author contributions: D.P. and R.J. jointly conceived the study, guided the analysis, and interpreted the results; M.K., D.P., and R.J. designed research; M.K. performed research; M.K. contributed new reagents/analytic tools; M.K. analyzed data; and M.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. Email: minkyoung.kimm@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1801429116/-/DCSupplemental.

Published online December 26, 2018.

www.pnas.org/cgi/doi/10.1073/pnas.1801429116

preceding outbreaks via intergroup dynamics, while abrupt growth but sharp decline in remote or peripheral regions is driven by rapid but inconstant feedback (abrupt outbreaks during an intensive period) via intragroup dynamics. Additionally, similar diffusion trends between two populous cities reflect synchronous feedback mechanisms of regional social systems, which is likely due to large volumes of external visitors and heavy reciprocal fluxes between the cities. That is, human mobility is a vital factor of mutual excitations across regions.

## Methods

We first explain our data collection and propose a diffusion framework. It quantifies exogenous and endogenous dynamics in disease spread and infers probabilistic transmission routes and cross-regional infection flow across a heterogeneous social system.

**Data Collection.** We investigate dengue outbreaks in Queensland, Australia, from 2002 to 2016, provided by Queensland Health. Dengue is a mosquito-borne viral disease transmitted among humans by mosquito vectors, whose outbreak risk is rapidly increasing worldwide (17). The data contains records of anonymized infected individuals such as onset dates, residence postcodes, and acquisition places if available. For understanding cross-regional infections at a macro level, we categorize residence postcodes into 15 regions, which correspond to the statistical areal level 4 defined by the Australian Statistical Geography Standard. Based on selected target regions, we create an event sequence of infections as a tuple consisting of occurrence time and region identity.

We also incorporate human mobility as topological heterogeneity across multiple regions, which reflects macrolevel internal dynamics in a social system. To obtain structural connectivity between regions, we use three different types of travel data such as International Visitor Survey, National Visitor Survey, and geo-tagged Twitter posts (see *SI Appendix*, section S1 for detailed statistics and measurements).

**Background.** We consider a Hawkes process (18) as our fundamental diffusion framework, since it is a non-Markovian extension of a Poisson process and thus realizes the clustering of events in the real world. A general univariate Hawkes process is defined with an intensity function,

$$\lambda(t) = \mu + \sum_{t_i < t} g(t - t_i), \qquad [1]$$

where the first term, $\mu$, represents the background intensity by external influence. The second term characterizes the endogenous feedback by weighting the influences of past events on future events. That is, the intensity of event occurrences is dependent on the history of preceding events.

Fig. 1 shows the embodiment of a Hawkes process to a disease-spread scenario. As shown in Fig. 1*A*, disease infections are represented as a single arrival process. It is reframed as Fig. 1*B* by considering self-excitations and mutual excitations (intraregion and interregion disease transmissions). As discussed in the Introduction, such cross-regional outbreaks are accelerated by human mobility (solid and dashed arrows in Fig. 1*C*). The infection pathways from regions where the vector is absent to other regions (dashed lines) result from an infected individual (international or domestic visitor) transiting through the vector-free regions. That is, human mobility allows bidirectional infection pathways (mutual excitations) among vector-free and vector-present regions.

In this context, the objective of our framework is to model bursty behavior (clustered in time and space) of disease outbreaks across metapopulations by incorporating human mobility as topological pathways in a heterogeneous social system.

**Proposed Model.** We now propose the LIPP model, which incorporates the exogeneity and endogeneity of a social system as major components for realizing the bursty diffusion processes in the real world. Based on inputs of a spatial and temporal event sequence and cross-regional human mobility, our model aims to quantify the reflexivity (level of feedback on prior events) (15, 16) of a social system using estimated model parameters and to infer transmission routes and infection flow between regions.

Suppose that we observe an event sequence ***D*** consisting of $N$ spatiotemporal events in a set of regions ***R*** during an observation time period $[0, T]$. Here, each event is represented by a pair of its occurrence time $0 < t_n < T$

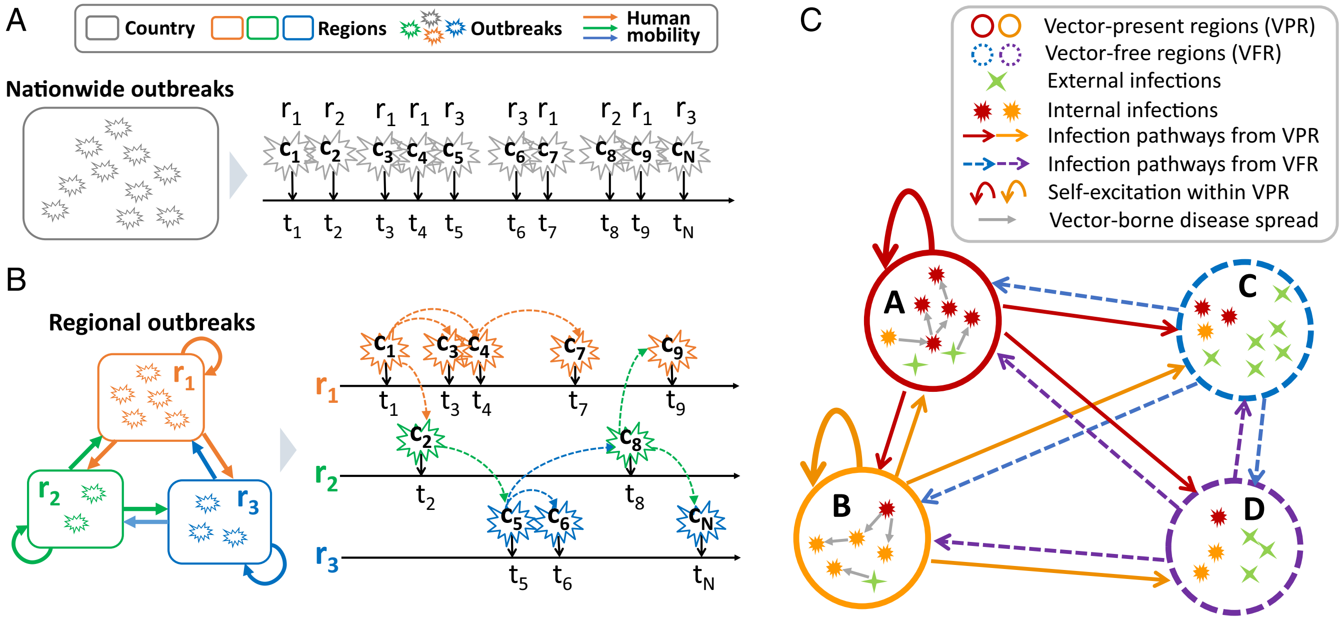
A
Country
Regions
Outbreaks
Human mobility
Nationwide outbreaks
r1 r2 r1 r1 r3 r3 r1 r2 r1 r3
c1 c2 c3 c4 c5 c6 c7 c8 c9 cN
t1 t2 t3 t4 t5 t6 t7 t8 t9 tN
B
Regional outbreaks
r1 r2 r3
c1 c3 c4 c7 c9 c2 c8 c5 c6 cN
t1 t3 t4 t7 t9 t2 t8 t5 t6 tN
C
Vector-present regions (VPR)
Vector-free regions (VFR)
External infections
Internal infections
Infection pathways from VPR
Infection pathways from VFR
Self-excitation within VPR
Vector-borne disease spread
A
B
C
D

**Fig. 1.** Process of disease spread. (*A*) Nationwide outbreaks of an infectious disease over time (*i*th contagion $c_i$ at time $t_i$ in region $r_i$, $N$ total outbreaks during an observation period). (*B*) Nationwide outbreaks in *A* can be decomposed into different timelines of each region, which can be represented as spatial and temporal point processes. Regions are color-coded, thick colored arrows represent human mobility between regions, and dashed arrows indicate hidden infection trajectories via social interactions. (*C*) Cross-regional infection pathways via human mobility. Intraregion infections (self-loops) occur in vector-present regions (solid circle), which are color-coded by their original regions (red and orange spikes). Interregion infections (arrows) are activated by human mobility both from vector-free (circles C and D) and vector-present (circles A and B) regions (VFR and VPR) to others (dashed and solid arrows). Vector-free regions (dashed circle) are contaminated by transitions of infected individuals, not by self-excitations (no self-loops in circles C and D). Both VPR and VFR receive infected individuals both from outside (e.g., international travelers, green crosses) and inside (e.g., domestic visitors, red and orange spikes) of a social system.

and region $r_n \in \boldsymbol{R}$ as a tuple such that $\boldsymbol{D} = \{(t_n, r_n)\}_{n=1}^{N}$, and the events are sorted by their time moments. As shown in Fig. 1*B*, we consider multiple timelines separated by event occurrence regions. For each region $r$, the history of events consists of two different types of event sequences, $\boldsymbol{D}^{r0}$ and $\boldsymbol{D}^{rk}$, influenced by external sources $0 \notin \boldsymbol{R}$ and triggered by preceding events at neighboring regions $k \in \boldsymbol{R}$, respectively. Given the whole event sequence $\boldsymbol{D}^{r} = \cup_{k \in \boldsymbol{R}_+} \boldsymbol{D}^{rk}$, $\boldsymbol{R}_+ \equiv \boldsymbol{R} \cup \{0\}$, we assume that each event sequence is generated by a Poisson process. Thus, the intensity function of region $r$ at time $t$, $\lambda_r(t)$, is defined based on the superposition property of Poisson processes (19) as

$$\lambda_r(t) = \lambda_r^0 + \sum_{k \in \boldsymbol{R}} \lambda_r^k(t). \qquad [2]$$

That is, we consider doubly stochastic point processes defined by $\lambda_r^k(t)$ as our diffusion framework for the realization of intraregion ($r = k$) and interregion ($r \neq k$) disease transmissions, which corresponds to a multidimensional Hawkes process.

We incorporate three major counterbalancing components into our framework: (*i*) exogenous influence covering environmental heterogeneity of target regions, (*ii*) endogenous influence attributed to macrolevel interactions between metapopulations, and (*iii*) a time decay effect with an exponential memory kernel. Details are discussed in *Exogeneity* and *Endogeneity*.

***Exogeneity.*** In region $r$, events can occur independently of a previous event history, due to external influence. This is modeled with a Poisson process with a background intensity,

$$\lambda_r^0 = \eta_r \, \rho_r^0, \qquad [3]$$

where $\eta_r > 0$ denotes disease-specific environmental heterogeneity of region $r$ (environmental infectiousness of a target disease). That is, region $r$ has an intrinsic environmental risk that does not change much over time, such as average temperature and humidity, annual precipitation, and distributions of disease vectors (e.g., mosquito vector for dengue virus), and thus some areas are more likely to experience disease outbreaks than others. The second term, $\rho_r^0 > 0$, represents the probability that an infection occurs in region $r$ by external exposures such that $\sum_{r \in \boldsymbol{R}} \rho_r^0 = 1$. For instance, international visitors from virus endemic regions outside $\boldsymbol{R}$ (i.e., region 0) trigger local outbreaks in region $r$.

***Endogeneity.*** Contrary to exogenous infections, internal dynamics in a social system drives bursts of events through interactions between individuals over social networks, so it is called internal influence (10). Our model incorporates cross-regional human mobility as macrolevel endogenous effects on diffusion, and the intensity brought by mutual excitations across multiple regions is defined as

$$\lambda_r^k(t) = \sum_{t_i < t} \zeta(r, k) \, \xi_k \rho_r^k \, \phi_r(t - t_i), \qquad [4]$$

where $\zeta(r, k) = \delta_{rk} \, (I(r \in \boldsymbol{V}) - 1) + 1$. Here, $\delta_{rk}$ is the Kronecker delta function, and $I(r \in \boldsymbol{V})$ denotes the indicator function verifying $r$ is in a set of vector-present regions $\boldsymbol{V}$. This $\zeta(\cdot, \cdot)$ function helps avoid self-excitations within vector-free regions, as discussed in Fig. 1*C*. The second term, $\xi_k > 0$, represents the latent influence (infectiousness) of region $k$ on other regions, due, in part, to population density, social interactivity, transportation hub, and vicinity to virus-endemic regions, embedding socioeconomic factors. The third term, $\rho_r^k > 0$, represents the strength of directed connectivity from region $k$ to $r$ based on human mobility patterns such that $\sum_{r \in \boldsymbol{R}} \rho_r^k = 1$. For instance, the center of a city has a larger floating population than neighboring suburbs (connected by commute or travel routes), and thus it more likely triggers further infections in its neighboring regions. That is, the latent influence ($\xi_k$) of region $k$ weights the human mobility $\rho_r^k$ from region $k$ to $r$. Finally, the last term, $\phi_r(\cdot)$, captures the time relaxation function for reflecting the effect of time decay on the likelihood of diffusion. For this aging effect, we consider an exponential memory kernel such that $\phi_r(t - t_i) = \exp(-\varphi_r(t - t_i))$, where $\varphi_r > 0$ indicates the time decay parameter for region $r$, i.e., the level of infectiousness decay in region $r$.

**Bayesian Inference.** We apply Bayesian inference to estimate the latent influence of each region in our proposed model by using a gamma distribution as a conjugate prior and thus obtaining a gamma posterior. We also introduce the latent index variables $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ consisting of event indicators each of which has triggered the $i$th event, since infection pathways are unknown as discussed in Fig. 1*B* (dashed lines). By using the stochastic expectation–maximization (EM) algorithm, we learn our model parameters and estimate probabilistic transmission routes (see *SI Appendix*, section S2 for details).

***Granger causality.*** A multivariate Hawkes process is a linear dependence structure of mutually exciting point processes, whose notion has been shown to reflect the Granger causality (20, 21). Granger causality is a statistical belief that the knowledge of a possible cause should improve ("Granger-causes") the prediction of the subsequent effect (22). Finding causality between two stochastic random variables in a time series is related to learning multivariate Hawkes kernels in parametric (4, 11, 23) or nonparametric ways (24, 25).

***Granger causality of multivariate Hawkes process.*** In the context of Granger causality, Eqs. **2**–**4** can be reformulated with a more general linear dependence structure of an $R$-dimensional Hawkes process as

$$\lambda_r(t) = \mu_r + \sum_{k=1}^{R} \int_0^t g_r^k(t - s) \, dN_k(s), \qquad [5]$$

where $g_r^k$ denotes the memory kernel for time decay effect of past type-$k$ events (events occurred in region $k$ in our case) at $s$ on type-$r$ events (events in region $r$) at $t$, and $N_k(s)$ is the number of type-$k$ events up to time $s$. Suppose the Granger causality graph $G = (\boldsymbol{R}, \boldsymbol{E})$, consisting of event types $\{k, r\} \in \boldsymbol{R}$ and directed causation edges $k \rightarrow r \in \boldsymbol{E}$ when type-$k$ events Granger-cause type-$r$ events. That is, the Hawkes memory kernel $g_r^k$ enables us to construct the Granger causality graph (21), i.e., cross-regional infection flow.

## Simulation

**Synthetic Data Generation.** We generate synthetic data by using an exact sampler (26), which samples time moments without approximation by decomposing a random variable into multidimensional nonhomogeneous Poisson processes based on the superposition property. This enables us to obtain marks (event types: regions) of triggering events to be used as ground truth of infection flow. Accordingly, Fig. 2 illustrates examples of true and estimated region-to-region matrices of infection flow from our synthetic datasets by varying the number of regions such that $|\boldsymbol{R}| \in \{3, 5, 10, 15\}$. For 120 event sequences, the average accuracy rate of our proposed model is greater than 85% (see *SI Appendix*, section S3 for our simulation algorithm).

**Robustness Test.** In real situations, infection reports are often missing or delayed, which makes it more challenging to learn model parameters. In this regard, we add variations to the synthetic data in three different ways: (*i*) random missing, (*ii*) clustered missing, and (*iii*) time delay by 5%, 10%, and 15% for each case (1,200 test cases in total). With the synthetic data, we conduct robustness tests with respect to the recovery of infection

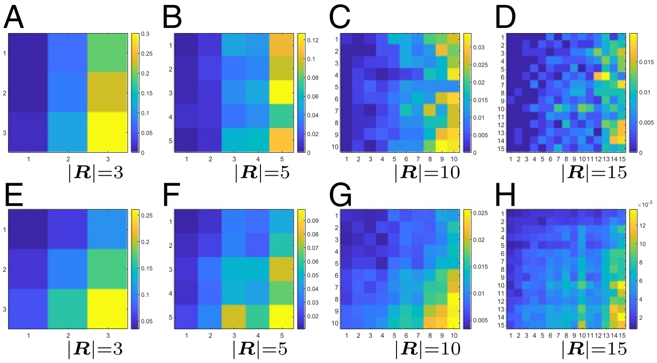

**Fig. 2.** Examples of (*A*–*D*) true and (*E*–*H*) estimated infection flow from synthetic datasets by varying the dimension of a multivariate Hawkes process. Estimations are conducted with a 95% confidence interval based on the inferred latent indicator variables. See *SI Appendix*, Fig. S6 for an additional comparison with baselines.

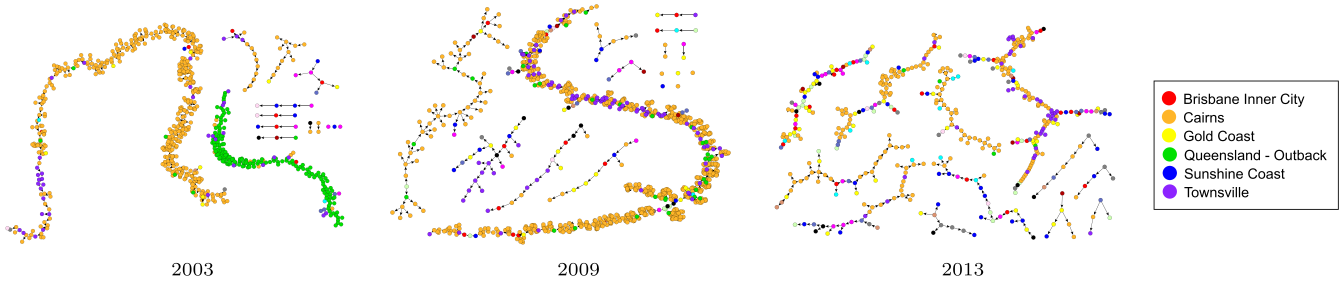

**Fig. 3.** Examples of constructed transmission routes between infections based on estimated pairwise probabilities of triggering and triggered dengue cases, for the selected years (2003, 2009, and 2013) with the largest dengue outbreaks. Each node indicates a dengue case, whose incoming and outgoing links are incident to preceding and triggered cases with the highest probabilities, for simplicity. Nodes are color-coded by regions, and only six region names with the largest outbreaks are presented, for brevity.

flow and model parameters. For the parameter recovery, we also evaluate relative strengths between estimated parameters, which is important to validate our subsequent interpretations of underlying diffusion processes with real data. As a result, our proposed model is robust to noisy data, as shown in *SI Appendix*, Table S3. Clustered missing data affect the model performance the most, followed by random missing and time-delayed data, but the accuracy rates remain over 70% (see *SI Appendix*, section S3 for data variations, and see *SI Appendix*, sections S4 and S4 for test results).

**Comparisons with Baselines.** We also compare our proposed model, "LIPP with prior" based on Bayesian inference with two baselines: (*i*) "LIPP without prior" based on the maximum likelihood estimation (MLE) and (*ii*) a recent competing approach, called "MLE-SGLP," which learns causality structures of nonparametric multivariate Hawkes processes based on MLE with sparse-group lasso (25). As evaluation metrics, inference errors are measured for all 1,200 synthetic datasets, which demonstrates that our model outperforms the two baselines (see *SI Appendix*, section S4 for details).

## Case Study: Dengue Spread

In this section, we conduct experiments on real data, whose results are interpreted with estimated model parameters based on the verification of parameter and infection-flow recovery with synthetic data in *Simulation*. For the experiments on real data, we set the observation time window as 1 y to examine time-evolving diffusion dynamics with a fine-grained time resolution.

**Cross-Regional Infection Flow.** As discussed earlier, infection pathways are unobservable, so we estimate the probability that each preceding event has triggered a current event by using the stochastic EM algorithm. Fig. 3 shows the examples of constructed transmission routes based on estimated pairwise probabilities of triggering and triggered dengue cases, for the 3 y with the largest outbreaks during a 15-y period. Here, each node presents a dengue case, color-coded by region. As the figure shows, earlier dengue outbreaks tend to be more locally clustered, but, over the years, they become globally interconnected across regions, leading to more complex behavior of infectious disease spread. Based on the estimated transmission routes in Fig. 3, the corresponding infection flow between regions are illustrated at a macro level in Fig. 4. As the figure shows, spread of dengue becomes more far-reaching across Queensland over time.

These all are consistent with event raster plots in Fig. 5, exhibiting increasing dengue outbreaks all over the regions throughout the year in 2013, compared with local outbreaks during the intensive period in 2003 and 2009. Such spatial expansion of infectious diseases can be attributed to the increase in travel volumes (12, 27).

**Reflexivity of a Regional Social System in Disease Spread.** A Hawkes process generalizes a nonhomogeneous Poisson process by allowing the self-exciting nature via preceding events, as discussed in Eq. **1**. The linearity of the conditional intensity $\lambda(t)$ helps quantify the level of exogeneity and endogeneity in diffusion processes and align with a branching process

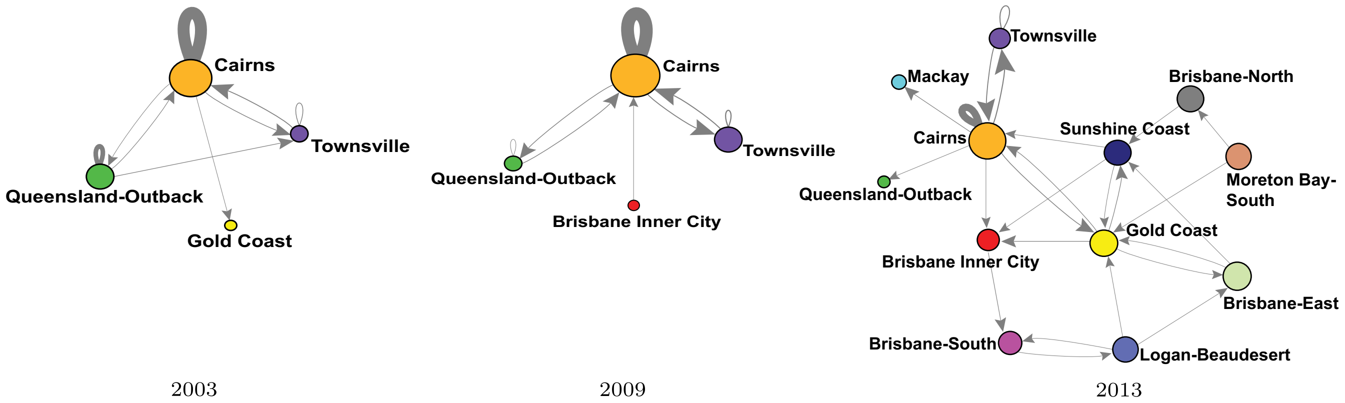

**Fig. 4.** Cross-regional infection flow in accordance with Fig. 3. Node are color-coded as in Fig. 3, labels denote region names in Queensland, and node sizes are proportional to the number of dengue cases that occurred. Arrow heads face influenced regions, and the width of a causal link indicates the strength of influence. Only links triggering more than 1% of dengue cases are presented, for brevity. Self-loops and links represent self-excitation and mutual excitation, respectively.

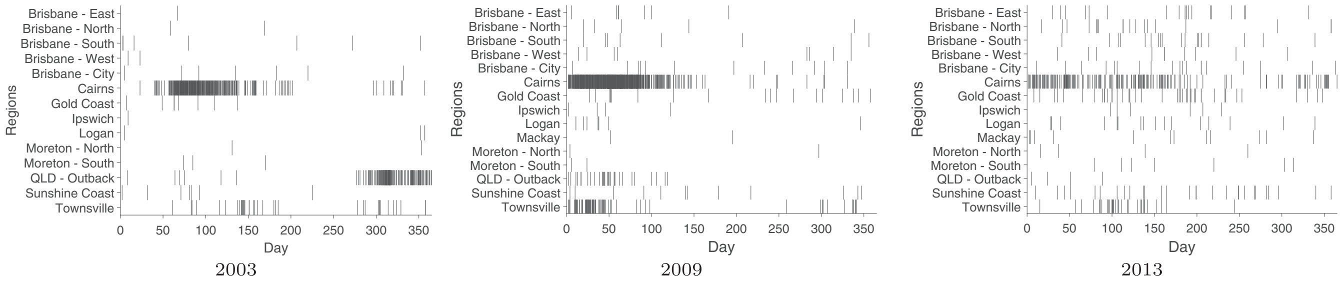

**Fig. 5.** Raster plots of event occurrences (dengue cases) during a year (*x* axis) in each region (*y* axis) for the selected years as in Figs. 3 and 4.

consisting of triggers and their descendants (28). The branching ratio $b$ represents the average number of triggered events per initiating event and is defined as $b = \int_0^\infty g(t)\,dt$ (15, 16). In our framework of an extended multivariate Hawkes process, the background intensity and branching ratio correspond to $\mu_r = \eta_r \rho_r^0$ and $b_r = \sum_{k \in \boldsymbol{R}} \frac{1}{\varphi_r} \zeta(r,k) \xi_k \rho_r^k$ based on each regional intensity $\lambda_r(t)$ in Eq. **2**. Accordingly, we quantify the level of exogeneity $\boldsymbol{\mu} = \{\mu_r\}_{r \in \boldsymbol{R}}$ and endogeneity $\boldsymbol{b} = \{b_r\}_{r \in \boldsymbol{R}}$ for all target regions based on the estimated parameter values with our real data.

***Behavioral split.*** Fig. 6 summarizes these quantifications for regions with the largest number of dengue cases during the 15-y period. In general, the background intensity $\boldsymbol{\mu}$ hardly changes, while the branching ratio $\boldsymbol{b}$ increases over time in metropolitan or populated areas such as Brisbane Inner City (BIC), Gold Coast (GC), and Sunshine Coast (SC) relative to remote or peripheral areas such as Cairns, Outback, and Townsville. In , *Left*, these two groups of regions also exhibit different growth patterns of dengue cases: precursory growth and symmetric decline in populous regions (BIC, GC, and SC) versus abrupt rise and sharp drop in peripheral regions (Cairns, Outback, and Townsville) in 2003 and 2009, showing a split in behavior. Additionally, mosquito vectors are presented in the three peripheral regions, while they are absent in the other populous areas.

***Intragroup and intergroup dynamics.*** Precursory growth in the major population centers is likely due to high reachability from statewide regions, i.e., high probability of importation of infected individuals. However, the absence of mosquito vectors in these regions allows no more excitations by previous outbreaks, leading to symmetric decline. In other words, dengue outbreaks in these populous regions are driven by strong but unsustainable intergroup dynamics. On the other hand, abrupt growth but sharp decline in the peripheral regions is attributed to rapid but inconstant excitations via mosquito vector transmissions. This strong but unstable intragroup dynamics is possibly affected by time-varying vector density and visitor volumes. That is, nonuniformly distributed mosquito vectors statewide, unbalanced human mobility between regions, and time-varying visitor volumes compositely lead to such behavioral split. Interestingly, BIC and GC exhibit similar growth patterns and reflexivity of a regional social system, which implies that endogenous feedback mechanisms are synchronous. This is likely due to human mobility patterns: The large volumes of external visitors and heavy

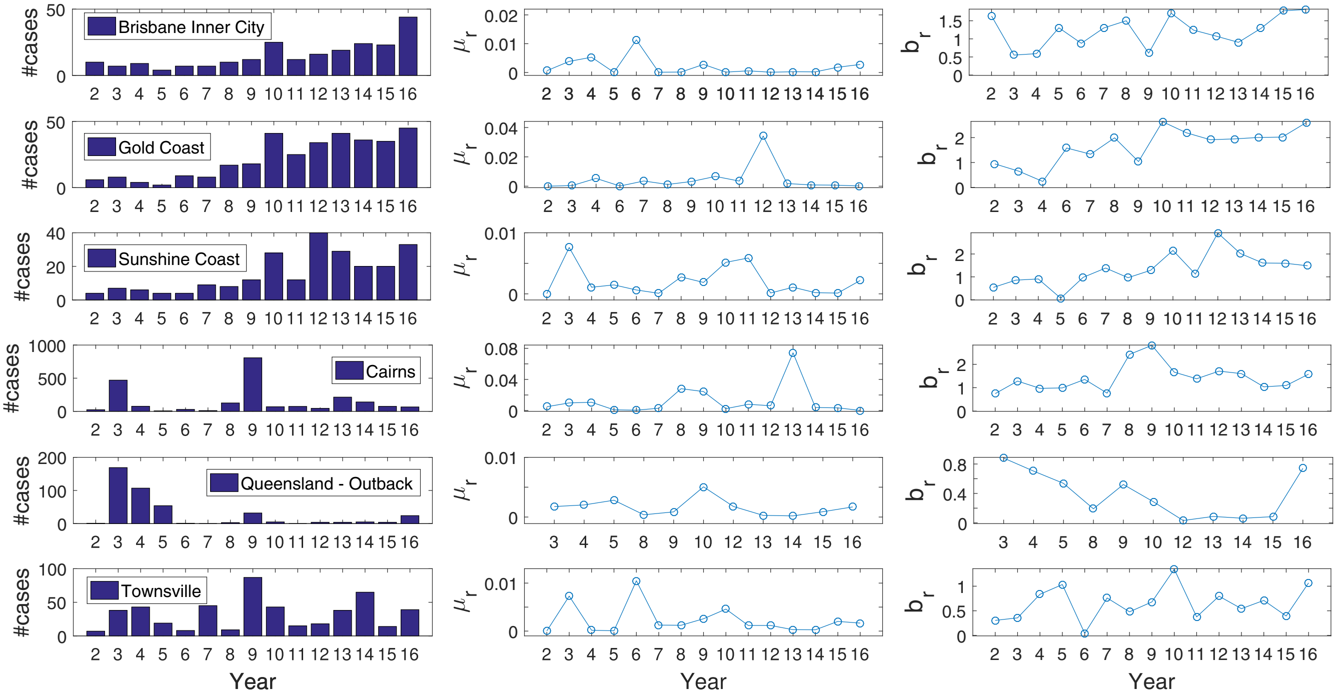

**Fig. 6.** Reflexivity of regional social systems. The plots show (*Left*) the distributions of dengue outbreaks in each region during our observation period, and the level of (*Middle*) exogeneity $\mu$ and (*Right*) endogeneity $b$ in dengue spread over time. Six regions are selectively chosen, with the largest number of cases among 15 regions. For brevity, years are presented with the last two digits of the 2000s.

reciprocal fluxes between the two cities more likely drive mutual excitations.

## Discussion

The spread of infectious diseases leads to formation of event clusters in both space and time. Such spatiotemporal events are well realized by a point process, due to its flexible consideration of lasting impact of bursty behaviors rather than a current snapshot (4), and thus it is widely used as a mathematical tool in diverse research areas (28, 29). In this context, we proposed a model, LIPP, which generalizes a multidimensional Hawkes process by incorporating macrolevel internal dynamics of metapopulations, driven by human mobility.

**Extension of Networked Hawkes Processes.** Our proposed memory kernel in Eq. **4** can be reformulated with element-wise matrix multiplication as

$$\lambda_r^k(t) = \boldsymbol{A}_{rk} \boldsymbol{W}_{rk} \boldsymbol{C}_{rk} \boldsymbol{\Phi}_{\boldsymbol{rk}}(\Delta t), \qquad \textbf{[6]}$$

where $\boldsymbol{A} \in \{0,1\}^{R\times R}$ is an adjacency matrix for connectivity across $R = |\boldsymbol{R}|$ regions, $\boldsymbol{W} \in \mathbb{R}_+^{R\times R}$ is a nonnegative weight matrix for human mobility weighted by latent influence of regions, $\boldsymbol{C} \in \{0,1\}^{R\times R}$ is a constraint matrix for avoiding self-excitations in vector-free regions, and $\boldsymbol{\Phi}(\Delta t) \in \mathbb{R}_+^{R\times R}$ is a time relaxation matrix for applying time decay effect. Parametric Hawkes models (4, 11, 23) learn the weighting scheme $\boldsymbol{W}$, but they embed $R \times R$ latent factors across different application domains, leading to more complex interpretation. Our model uses domain knowledge ($R \times R$ human mobility) to decouple $R$ latent influence from the weighting scheme $\boldsymbol{W}$. In this way, we can reduce the complexity and focus more on the hidden nature of each region, $r \in \boldsymbol{R}$. By introducing a constraint matrix $\boldsymbol{C}$, the proposed model covers vector-borne (indirect) and contact-based (direct) diffusion processes. By introducing latent indicator variables for triggering events, we can obtain transmission routes between events and cross-regional infection flow, whereas prior work has largely focused on dependency structure between Hawkes kernels (21, 25).

**Cross-Domain Implications.** In real situations, tracking infection routes often depends on time-consuming site investigations or a survey on travel routes of infected patients. Based on such efforts and expert knowledge, a single outbreak identification (ID) is assigned to a collection of cascading (or ongoing) local transmission possibly initiated by the same index case (see *SI Appendix*, Fig. S2 for the reference of outbreak IDs provided by Queensland Health). Outbreak IDs are currently the best-known data source for coupling cases, but a considerable proportion of cases are left unknown or possibly misidentified, without linkages between coupled cases. Here, our estimation of probabilistic transmission routes can provide investigators or experts with initial reference of infection pathways for their efficient tracking and timely control of disease spread, reducing response time and cost under resource constraints.

For understanding the origin of a burst, the interplay between external shock and internal dynamics in complex systems has also been of great interest across disciplines (10, 30). We quantified the level of exogeneity and endogenity of clustered bursts by incorporating environmental heterogeneity and internal dynamics between metapopulations. That is, our approach can reveal rich context which underlies time-evolving subgroup interactions in the real world.

All these aspects increase the applicability of our proposed model to a wide range of intragroup and intergroup diffusion processes in social and natural systems at a macro level. Additionally, microlevel investigations, such as targeting subregions and analyzing detailed socioeconomic factors, would help obtain a holistic view of underlying diffusion mechanisms, which is an interesting direction for future work.

**ACKNOWLEDGMENTS.** We would like to thank Queensland Health and Tourism Research Australia for providing data. Also, we thank Cassie Jansen at Queensland Health for the key role in guiding us in defining the problem that needs to be solved.

1. Barabasi A (2005) The origin of bursts and heavy tails in human dynamics. *Nature* 435:207–211.
2. Kim M, Newth D, Christen P (2013) Modeling dynamics of meta-populations with a probabilistic approach: Global diffusion in social media. *CIKM'13* (Assoc Comput Machinery, New York), pp 489–498.
3. Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J (2015) SEISMIC. *KDD'15* (Assoc Comput Machinery, New York), pp 1513–1522.
4. Kim M, McFarland D, Leskovec J (2017) Modeling affinity based popularity dynamics. *CIKM'17* (Assoc Comput Machinery, New York), pp 477–486.
5. Shen H, Wang D, Song C, Barabási AL (2014) Modeling and predicting popularity dynamics via reinforced Poisson processes. *AAAI'14* (Assoc Advancement Artificial Intelligence, Menlo Park, CA), pp 291–297.
6. Mohler GO, Short MB, Brantingham PJ, Schoenberg FP, Tita GE (2011) Self-exciting point process modeling of crime. *J Am Stat Assoc* 106:100–108.
7. Short MB, Bertozzi AL, Brantingham PJ (2010) Nonlinear patterns in urban crime: Hotspots, bifurcations, and suppression. *SIAM J Appl Dyn Syst* 9:462–483.
8. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402: 605–609.
9. Crane R, Sornette D (2008) Robust dynamic classes revealed by measuring the response function of a social system. *Proc Natl Acad Sci USA* 105:15649–15653.
10. Kim M, Paini D, Jurdak R (2018) Real-world diffusion dynamics based on point process approaches: A review. *Artif Intell Rev* 28:1–30.
11. Iwata T, Shah A, Ghahramani Z (2013) Discovering Latent Influence in Online Social Activities via Shared Cascade Poisson Processes. *KDD'13* (Assoc Comput Machinery, New York), pp 266–274.
12. Bhatt S, et al. (2013) The global distribution and burden of dengue. *Nature* 496: 504–507.
13. Colizza V, Vespignani A (2008) Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *J Theo Bio* 251:450–467.
14. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A (2015) Epidemic processes in complex networks. *Rev Mod Phys* 87:925.
15. Filimonov V, Sornette D (2012) Quantifying reflexivity in financial markets: Toward a prediction of flash crashes. *Phys Rev E* 85:056108.
16. Hardiman SJ, Bercot N, Bouchaud JP (2013) Critical reflexivity in financial markets: A Hawkes process analysis. *Eur Phys J B* 86(10):442.
17. World Health Organization (2009) *Dengue: Guidelines for Diagnosis, Treatment, Prevention and Control* (World Health Org, Geneva).
18. Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58:83–90.
19. Cinlar E (2013) *Introduction to Stochastic Processes* (Courier Corp, Chelmsford, MA).
20. Didelez V (2008) Graphical models for marked point processes based on local independence. *J R Stat Soc Ser B Stat Methodol* 70:245–264.
21. Eichler M, Dahlhaus R, Dueck J (2017) Graphical modeling for multivariate hawkes processes with nonparametric link functions. *J Time Ser Anal* 38:225–242.
22. Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica J Econometric Soc* 37:424–438.
23. Linderman S, Adams R (2014) Discovering latent network structure in point process data. *J Machine Learning Res* 32:1413–1421.
24. Lewis E, Mohler G (2011) A nonparametric em algorithm for multiscale Hawkes processes. *J Nonparametric Stat* 1:1–20.
25. Xu H, Farajtabar M, Zha H (2016) Learning Granger causality for Hawkes processes. *Proc Machine Learning Res* 48:1717–1726.
26. Dassios A, Zhao H (2013) Exact simulation of Hawkes process with exponentially decaying intensity. *Electron Commun Probab* 18:1–13.
27. Wesolowski A, et al. (2015) Impact of human mobility on the emergence of dengue epidemics in Pakistan. *Proc Natl Acad Sci USA* 112:11887–11892.
28. Daley DJ, Vere-Jones D (2007) *An Introduction to the Theory of Point Processes* (Springer, New York), Vol II.
29. Snyder DL, Miller MI (2012) *Random Point Processes in Time and Space* (Springer, New York).
30. Roehner B, Sornette D, Andersen JV (2004) Response functions to critical shocks in social sciences: An empirical and numerical study. *Intl J Mod Phys C* 15:809–834.