PNAS RESEARCH ARTICLE | ECOLOGY OPEN ACCESS

A simplicial epidemic model for COVID-19 spread analysis

Yuzhou Chen[a,1], Yulia R. Gel[b,c,1], Madhav V. Marathe[d,e,1], and H. Vincent Poor[f,1,2]

Contributed by H. Vincent Poor; received August 2, 2023; accepted November 11, 2023; reviewed by Maksim Kitsak and Shashanka Ubaru

Networks allow us to describe a wide range of interaction phenomena that occur in complex systems arising in such diverse fields of knowledge as neuroscience, engineering, ecology, finance, and social sciences. Until very recently, the primary focus of network models and tools has been on describing the pairwise relationships between system entities. However, increasingly more studies indicate that polyadic or higher-order group relationships among multiple network entities may be the key toward better understanding of the intrinsic mechanisms behind the functionality of complex systems. Such group interactions can be, in turn, described in a holistic manner by simplicial complexes of graphs. Inspired by these recently emerging results on the utility of the simplicial geometry of complex networks for contagion propagation and armed with a large-scale synthetic social contact network (also known as a digital twin) of the population in the U.S. state of Virginia, in this paper, we aim to glean insights into the role of higher-order social interactions and the associated varying social group determinants on COVID-19 propagation and mitigation measures.

digital twin | synthetic social contact network | COVID-19 | forecasting disease dynamics

Significance

Recent results demonstrate that higher-order relationships among entities of complex systems play an important role behind system functionality, which is in contrast to the focus of more conventional studies that target pairwise interactions. Inspired by the recently emerged ideas on the latent relationships between the viral dose received by the individual through social interactions and risk of infection referred to as inoculum, we study how higher-order, or group interactions in social networks such as family, religious, and shopping activities, described by the notion of simplicial complexes, impact dynamics of disease transmission and the associated mitigation strategies. The significance of the proposed ideas is illustrated through analysis of COVID-19 transmission over large synthetic social proximity network of the State of Virginia.

Complex networks provide us a versatile machinery of methods to elucidate and systematize a wide variety of disparate phenomena arising anywhere from social communications to human brain connectome to power grids to digital asset transactions. Until very recently, the tools of complex networks have predominantly focused on the description of dyadic, or pairwise interactions among nodes, for example, information propagation between two persons on a social media platform or transaction volume between two blockchain addresses. However, the emerging results in various domains of knowledge increasingly more often indicate that group interactions, that is, polyadic relationships among multiple constituents of the complex system, are the key toward better understanding the intrinsic mechanisms behind the system functionality (1–3). For instance, in ecology, two species may influence a focal species in an interactive manner, and as shown by Gibbs et al. (4), such higher-order relationships may bring a new light to modeling species coexistence, equilibrium dynamic, and the associated biodiversity of the ecosystem. In neuroscience, Herzog et al. (5) found that higher-order functional connectivity in brain networks may play the important complementary role in differentiating various neurodegenerative conditions such as dementia and Alzheimer's disease. As one could expect, money laundering schemes are inherently based on complex multi-entity relationships aiming to obfuscate fraudulent behavior, and identification of such recurrent transaction patterns is one of the primary approaches in deterring ransomware and other malicious activity in both the traditional finance and the ecosystem of digital assets (6). In turn, some recent results of social networks suggest that integration of polyadic relationships into the link prediction tools can noticeably enhance the algorithm performance, even in the case when the goal is only to predict the traditional dyadic links between two nodes (7). This phenomenon can be explained by the critical role that groups such as family, friends, and co-workers play in the message-passing mechanisms of social interactions. Not surprisingly, a similar phenomenon has been also documented in the analysis of contagion dynamics (8–11). Such important higher-order group interactions can be described in a holistic and mathematically rigorous manner by simplicial complexes (1, 12). The alternative, yet closely connected approaches to model the polyadic relationship is via hypergraphs (13–15) which, however, may arguably be viewed as somewhat less tractable, or via heuristics of network motifs (16–18). While offering important insights into the role of group interactions, these approaches for integrating higher-order properties into the analysis of disease propagation on social networks either tend to be restricted by considering a constant fixed transmission rate, thereby, limiting the impact of variability among social individual and group determinants or by focusing on simpler compartmental models on small-scale social networks (8, 12, 14, 19), thereby not explicitly accounting for the interplay between the impact of polyadic social interactions and the recovery rate.

Author contributions: Y.C., Y.R.G., M.V.M., and H.V.P. designed research; performed research; contributed new reagents/analytic tools; analyzed data; and wrote the paper.

Reviewers: M.K., Delft Institute of Technology; and S.U., International Business Machines (IBM) Research.

The authors declare no competing interest.

[1]Y.C., Y.R.G., M.V.M., and H.V.P. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: poor@princeton.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2313171120/-/DCSupplemental.

Published December 26, 2023.

Why Higher-Order Interactions Matter for Pathogen Spread? Over the last two decades, epidemiologists have considered the independent cascade models for between-host transmission. This modeling approach is attractive for several reasons, including its simplicity and mathematical tractability. An elegant generalization of this approach has been developed by Dodds and Watts (20), which includes the independent cascade model and complex contagion models (21) as special cases. Such a generalized contagion model incorporates three important components: i) the number of viral particles transmitted from person a to person b; ii) the total dose received by an individual due to multiple contacts, and iii) time within which this dose is received. However, all these models are *inherently linear* and do not take into account *higher-order interactions*. Nevertheless, it is important to note that the model of Dodds and Watts (20) leads to several insights regarding the minimum quantity of infectious particles required to establish infection, and as discussed and summarized by Van Damme (22), this minimum quantity is pathogen dependent. Remarkably, this phenomenon may result in a number of new important implications for modeling spread of the infectious diseases, in light of the recent results on the relationship between the dose received by an individual through social interactions and the risk of infection.

In particular, in a set of the most recent papers (23–25), authors point out that the risk of infection and the severity due to COVID-19 is likely to be dependent on *inoculum*: the viral dose received by an individual via interactions with other infected individuals that leads to an infection. In simpler terms, it means that if you and your neighbors are all together roughly at the same time then you are likely to get a higher dosage. The three components studied by Dodds and Watts (20) are relevant here. Experimental results discussed in refs. 22 and 23 point to the evidence that the probability of an individual getting infected is higher in large gatherings. One way, of course, to capture the impact of large gatherings is to use the model of Dodds and Watt (20). An alternate way to describe the dosage due to multi-party interactions is to develop a model that captures the fact that a node u and its neighbors are all there at the same time. That would give us the triad or higher-order structures and the substructures, so each has a force of infection but the force is amplified by the simultaneous presence at the same location. Mathematically, such interaction substructures can be described by simplicial complexes. Hence, for instance, a triangle (i.e., a 2-simplex) would capture close proximity between more than three individuals simultaneously within a short period of time. As a result, we can model the concentration of viral particles in a given unit of time. A natural question would be then: *Can a simpler linear model capture the intended effect?* This does not appear to be the case. In other words, a fixed set of weights cannot capture these higher-order interactions. Let us consider a simple example, where nodes a, b, and c form a triangle. Our goal is to capture the force of infection of b and c on a. Let us use $\{0, 1\}$ to represent the state of the node (1 is infected and 0 otherwise). If both b and c are infected, the total force of infection due to b and c is proportional to 2 in a linear additive model. Using simplicial complexes, this would be proportional to 3. No assignment of weights edges on (a, b) and (a, c) can represent this additional force of infection. Moreover, we provide several toy examples in *SI Appendix*, section 2 to illustrate why higher-order interactions matter.

Inspired by these ideas and armed with the digital twin of the population in the state of Virginia, here we make the next step toward a fundamental understanding of the relationships between the level of dose and risk of infection. In particular, we combine the techniques from simplicial geometry of complex networks, epidemiology, and statistics, to shed more light on the following research four questions: (Q1) How can one combine in a holistic manner both the key individual and group characteristics in modeling contagion propagation? (Q2) How do the higher-order group interactions impact the transmission rate? (Q3) How can the transmission likelihood between an infected individual and a susceptible individual be inferred based on their varying social group determinants? (Q4) What role do the higher-order social interactions play in selecting intervention strategies?

1. Background on Simplicial Complex

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an attributed graph, where $\mathcal{V}$ is a set of nodes ($|\mathcal{V}| = N$) and $\mathcal{E}$ is a set of edges ($|\mathcal{E}| = M$). Let d_{uv} be the distance on $\mathcal{G}$ defined as the shortest path between nodes u and v, $u, v \in \mathcal{V}$, and $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be a symmetric adjacency matrix such that $\boldsymbol{A}_{uv} = \omega_{uv}$ if nodes u and v are connected and 0, otherwise (here, ω_{uv} is an edge weight and $\omega_{uv} \equiv 1$ for unweighted graphs). Furthermore, $\boldsymbol{D}$ represents the degree matrix with $\boldsymbol{D}_{uu} = \sum_v \boldsymbol{A}_{uv}$, corresponding to $\boldsymbol{A}$.

Definition 1.1: (Simplicial Complexes) Let $\mathcal{V}$ be a finite set of vertices. A *k-simplex* $\mathcal{S}^k$ is a subset of $\mathcal{V}$ of cardinality $k + 1$ (we do not allow $\mathcal{S}^k$ to be a multi-set, i.e., there are no repeated elements in $\mathcal{S}^k$). A *simplicial complex* (SC) $\mathcal{X}$ is a set of simplices with the property that if $\mathcal{S} \in \mathcal{X}$, then all subsets of $\mathcal{S}$ are also in $\mathcal{X}$.

Hence, nodes of $\mathcal{G}$ are 0-simplices, edges are 1-simplices, and triangles are 2-simplices. Fig. 1 shows a toy example of k-simplex, where $k = \{0, 1, 2, 3\}$. For a k-simplex of $k > 0$, we can also define its *orientation* by (arbitrary) selecting some order for its nodes, and two orderings are said to be equivalent if they differ by an even permutation. As a result, for a given k-simplex $\mathcal{S}^k$ with orientation $[i_0, i_2, \ldots, i_k]$, any face of $\mathcal{S}^k$ is assigned its own orientation (or "identifier") $[i_0, i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_k]$ (i.e., we omit the j-th element).

2. Simplicial Complex-Based Susceptible Infectious Recover (SIR)

To illustrate the role that polyadic interactions play in the propagation of the infectious agent on social networks, without loss of generality, we start by integrating the higher-order characteristics into the basic Susceptible Infectious Recover (SIR) model. (It is important to note that the proposed simplicial approach is not restricted to SIR and it can be combined with a general class of complex mechanistic models. We present more details on the potential generalizations in the subsequent sections.) In a classic SIR model (referred to as base SIR), the dynamics of the system at each timestamp can be described by the following system of equations, i.e.,

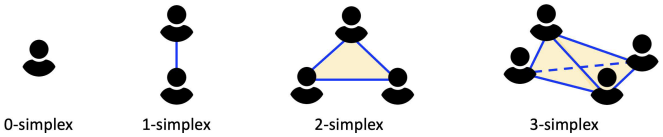

Fig. 1. A toy example of simplicial complexes in different dimensions.

$$\frac{dS}{dt} = -\frac{\beta SI}{N}, \frac{dI}{dt} = \frac{\beta SI}{N} - \gamma I, \frac{dR}{dt} = \gamma I,$$

where an individual can be in one of the three states: (S) susceptible, (I) infected and can infect susceptible nodes, and (R) recovered at any given time step t. Here, β and γ are the transition rate and recovery rate from S to I, and I to R, respectively. Unlike the above equation-based SIR model, in graph-based SIR models, we take into account the contact patterns between individuals by simulating the spread of a disease over a contact network. Specifically, each individual becomes a node in the network and the edges represent the associated connections between people (e.g., social contacts or physical proximity). In addition, similar to ref. 26, the model can be further expanded by adding a vaccinated state where the increment of the vaccinated pool depends on the effectiveness of the vaccine (e.g., the number of vaccine doses). We leave the exploration of the impact of vaccination and the comparison between a cumulative number of people exposed and the cumulative amount of vaccine for the spread of infection for future work.

In particular, we consider two graph-based SIR models, i.e., i) regular SIR and ii) simplicial complex-based SIR (called SC-SIR) for regular contact and family-based contact networks respectively. The main difference between the two SIR models is that the SC-SIR model considers additional/hidden infection transmission from high-dimensional simplicial complexes (e.g., filled triangles). Next, to integrate both the individual and group characteristics into the analysis of contagion propagation (i.e., our research Q1), we introduce different channels of infection for a susceptible node u in the SC-SIR model. That is, as shown in Fig. 2, the node u is in contact with one (Fig. 2 *A*–*D*) or more (Fig. 2*E*) infected nodes through links (1-simplices), and it becomes infected with probability β at each time step through each of these links. In Fig. 2 *E* and *F*, the node u is involved in a 2-simplex (triangle). In Fig. 2 *C* and *D*, one of the nodes of the 2-simplex is not infected, so node u can only receive the infection from the (red) link, with probability β. In Fig. 2*F*, the two other nodes of the 2-simplex are infected. Hence, node u can get the infection through both two 1-simplices (links) with probability β and the 2-simplex with additional probability $\Delta\beta$. Fig. 2*G* shows a filled triangle that represents all three nodes are infection status in the 2-simplex. Algorithm 1 outlines the pseudo-code for our proposed SC-SIR model (moreover, we also provide the pseudo-code of a variant of SC-SIR model, please refer to *SI Appendix*, section 7). Given a contact network $\mathcal{G}$, we assume that a node comes into contact with all its neighbours at each timestamp. For the sake of simplicity, we omit the timestamp t in the following discussion. More specifically, at each timestamp, the susceptible individual u will become infected from 2 contagion channels, i.e., i) 1-hop *infected* neighbors with a transmission probability $1-(1-\beta)^{N_I(u)}$ (here $N_I(u)$ denotes the number of 1-hop neighbors of the node u who are infected), and ii) *infected* higher-order structures (in this study we consider 2-simplices) which contain two infected individuals with a transmission probability $1-(1-\Delta\beta)^{Q_I(u)}$ (here $Q_I(u)$ denotes the number of 2-simplices which involve the susceptible node u and two other infected nodes). In our implementation, for the target susceptible node u at timestamp t, we first generate a random probability $\rho_{S,u}^{(t)}$ from a uniform distribution $U(0,1)$

Algorithm 1: SC-SIR

Input: Network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; transmission probability from a 0-simplex β, transmission probability from a 2-simplex $\Delta\beta$, recovery rate γ; S_0, I_0, R_0: the numbers of initially susceptible nodes, initially infected nodes, and initially recovered nodes, respectively.

Output: S_T, I_T, R_T: the numbers of susceptible nodes, infected nodes, and recovered nodes at time T respectively.

for $t = 1, 2, \cdots, T$ **do**
  for $i = 1, 2, \cdots, N$ **do**
    Compute $N_I^{(t)}(u_i)$ and $Q_I^{(t)}(u_i)$
    if $Status^{(t)}[u_i] == S$ **then**
      $\rho_{S,u_i}^{(t)} \sim U(0,1)$ ▷ Stochastic individual characteristics of node u associated at timestamp t
      if $\rho_{u_i}^{(t)} < 1-(1-\beta)^{N_I^{(t)}(u_i)}$ **then**
        $Status^{(t)}[u_i] = I$
      else if $\rho_{u_i}^{(t)} < 1-(1-\Delta\beta)^{Q_I^{(t)}(u_i)}$ **then**
        $Status^{(t)}[u_i] = I$
    else if $Status^{(t)}[u_i] == I$ **then**
      $\rho_{I,u_i}^{(t)} \sim U(0,1)$ ▷ Stochastic individual characteristics of the infected node u related to its recovery rate at timestamp t
      if $\rho_{u_i}^{(t)} < \gamma$ **then**
        $Status^{(t)}[u_i] == R$
  $S_t = \{Status^{(t)} == S\}$, $I_t = \{Status^{(t)} == I\}$, $R_t = \{Status^{(t)} == R\}$

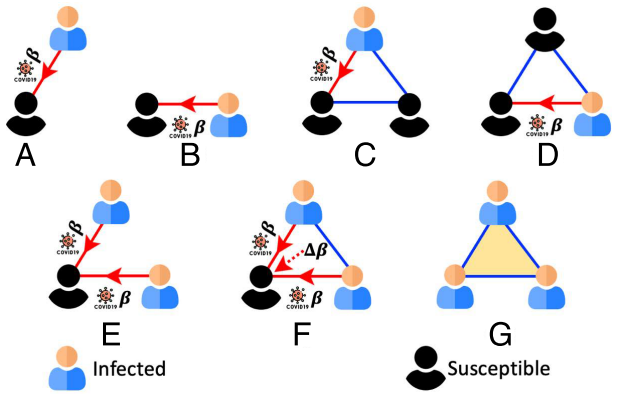

Fig. 2. Susceptible and infected nodes are colored in black and blue, respectively; (*A*–*G*) show different channels of infection for a susceptible node u (i.e., the left bottom node) in the SC-SIR model. See more detailed descriptions in Section 2.

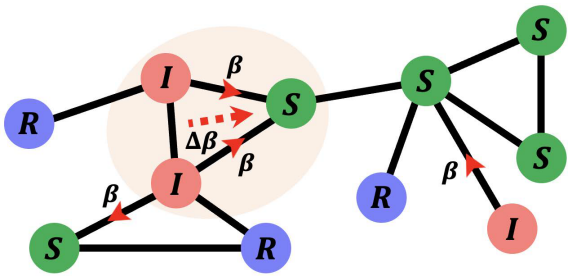

Fig. 3. Transmission of the COVID-19 virus over the contact network.

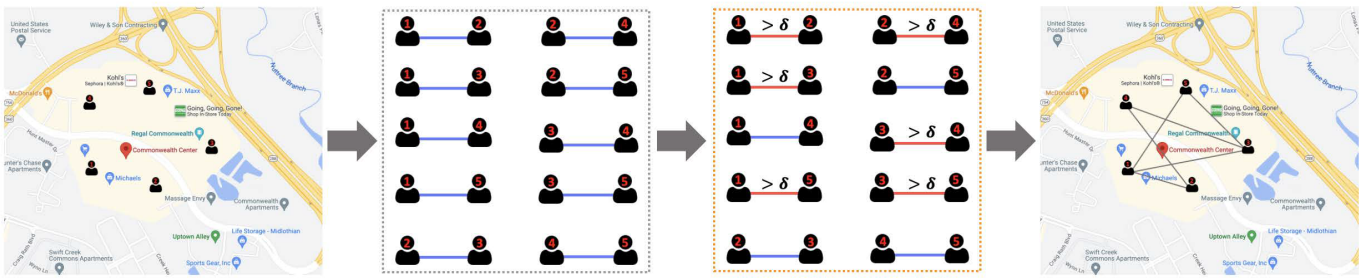

Fig. 4. Illustration of the activity-driven contact network construction in one activity location in one county, VA. In this toy example, we i) identify individuals active in a particular shopping mall on date t (for instance, we identified five such individuals), ii) compute the duration of contact between any two individuals, iii) establish the actual connections between two individuals based on the predefined threshold δ, and iv) build the corresponding contact network on the date t.

and compare it with the probability $1-(1-\beta)^{N_I(u)}$ of being infected by the 1-hop neighbors. If $\rho_{S,u}^{(t)} < 1-(1-\beta)^{N_I(u)}$, then the susceptible node u becomes infected. Otherwise, we compare $\rho_{S,u}^{(t)}$ with the probability $1-(1-\Delta\beta)^{Q_I(u)}$; similarly, if $\rho_{S,u}^{(t)} < 1-(1-\Delta\beta)^{Q_I(u)}$, then the susceptible node u becomes infected. (Here, $\rho_{S,u}^{(t)}$ may be viewed as stochastic individual characteristics of node u at timestamp t, while $\rho_{I,u}^{(t)}$ may be viewed as the individual characteristics of the infected node u related to a recovery rate at timestamp t.) To study diffusion among higher-order substructures of $\mathcal{G}$, we now form a real-valued vector space C^k which is endowed with basis from the oriented k-simplices and whose elements are called *k-chains*. Diffusion through higher-order graph substructures can be then defined via linear maps among spaces C^k of k-chains on $\mathcal{G}$. Fig. 3 visualizes how the infected nodes can infect their neighbouring nodes, changing their state from S to I. Note that, similarly to refs. 27–30, to capture the real-time evolution of the spread and to enhance the explainability of the obtained results, the above simplicial framework can be extended to a case of the time-varying model where transmission and recovery rates evolve over time (31).

Furthermore, to address our research Q3, in the proposed SC-SIR model, we infer the transmission rate between an infected individual u and a susceptible individual v based on their social determinants. In particular, given a contact network $\mathcal{G}$, we can extract demographic features $\boldsymbol{X} \in \mathbb{R}^{N\times F}$ of individuals including age, gender, and their household locations. In this case, we first compute the similarity sim_{uv} between nodes u and v. There are many ways to obtain sim_{uv}, and we list two options here, in which x_u and x_v are feature vectors of nodes u and v, i.e.,

- Cosine similarity. It uses the cosine value of the angle between two vectors to measure the similarity

$$sim_{uv} = \frac{x_u \cdot x_v}{|x_u||x_v|}. \quad [1]$$

- Heat Kernel. The similarity between two nodes u and v, where σ is the time parameter in heat conduction equation

$$sim_{uv} = \exp\left(-\frac{||x_u - x_v||_2}{\sigma}\right). \quad [2]$$

By definition, $\tilde{\beta}$ represents the *attribute-based* likelihood that a disease is transmitted from an infect node u to a susceptible node v per unit time. Given the pre-defined transmission rate β, we can infer the *attribute-based* transmission rate $\tilde{\beta}$ can be expressed as

$$\tilde{\beta}_{uv} = sim_{uv} \cdot \beta. \quad [3]$$

Hence, the susceptible individual u will become infected from 1-hop *infected* neighbors, where $N_I(u) = \{v_1, v_2, \ldots, v_{N_I(u)}\}$, with a transmission probability $1-\Pi_{i=1}^{N_I(u)}(1-\tilde{\beta}_{uv_i})$. Note that, we can fit the classic SIR model to real data to estimate β or we can select β via the cross-validation (see more details in *SI Appendix*, section 9) and we provide transmission rates of the SC-SIR model comparison in *SI Appendix*, section 6 and Table S5. See *SI Appendix*, section 1 and Table S1 for a full glossary.

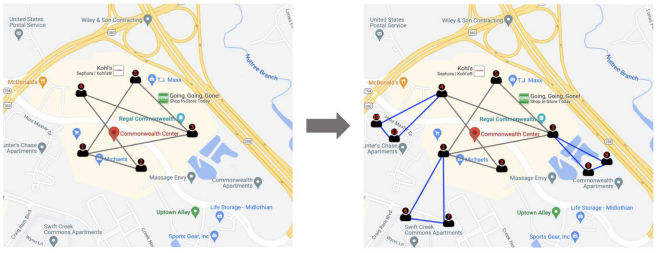

Fig. 5. Illustration of the activity-family-driven contact network construction. Specifically, equipped with the built the activity-driven contact network (from Fig. 4), we can build the resulting activity-family-driven contact network by adding additional connectivity information (i.e., connections between family members) for each individual based on their household information.

3. Experiments

A. Contact Network Construction. In this project, we study a synthetic social contact network of the U.S. state of Virginia. This dataset is characterized by the following summary of network statistics as follows, i.e., population—7,908,211, households—3,206,012, residence locations—3,206,012, and activity locations—729,228. Our goal is to forecast the number of positive COVID-19 cases at the county-level. For a target county, we first build the synthetic social contact network based on the longitude and latitude of households and activity locations (or we can use the ADCW ID* to select the target county). This is done to ensure that the targeted county is less impacted by external mobility as, for example, occurs in large metropolitan areas and university campuses. Now, we turn to the construction of a synthetic social contact network for the county c_i.

* https://www.adci.com/adc-worldmap.

Table 1. Summary of counties in the State of Virginia

Dataset	Population	Density	Diversity	Med. age	People fully vaccinated	Health ranking	Time range
Albemarle	113,535	157	0.379	39.4	87.47%	6/133	03/02/2022–03/31/2022
Charlotte	11,448	24	0.464	51.3	54.90%	118/133	12/22/2021–01/20/2022
Clarke	14,726	84	0.673	48.1	68.52%	19/133	02/10/2022–03/11/2022
Culpeper	53,596	141	0.452	39.8	65.24%	39/133	07/25/2021–08/23/2021
Hanover	111,603	236	0.269	42.7	76.96%	12/133	12/22/2021–01/20/2022
Highland	2,226	5	0.049	59.5	59.67%	66/133	12/22/2021–01/20/2022
Prince Edward	21,932	62	0.502	32.1	46.20%	92/133	12/22/2021–01/20/2022

S1 Using the latitude and longitude of each household (or the ADCW ID), we find the county where it is located. Hence, we can obtain household location IDs (hid^{c_i}), person IDs ($pid_w^{c_i}$; whose households are located in the county c_i), activity location IDs ($alid^{c_i}$) within the target county c_i. Moreover, in the "Activity Location Assignment" file, we also have timing information about the start time (i.e., the start time of activity in seconds since midnight Sunday/Monday) and during (i.e., duration of the activity in seconds) of each activity, and there are four types of activities—shopping, school, college, and religion.

S2 Based on the information obtained in the S1, for each activity location $alid_j^{c_i} \in alid^{c_i}$, we can extract i) individuals who are involved in an activity in the county c_i but do not live in the county c_i (denoted as $pid_{w/o}^{c_i}$), ii) total duration of a person (who can be either $pid_w^{c_i}$ or $pid_{w/o}^{c_i}$) in the activity location $alid_j^{c_i}$. For the sake of simplicity, we use pid^{c_i} to denote individuals in the county c_i.

S3 To build a synthetic social contact network of the county c_i based on its activity locations, we treat individuals pid^{c_i} as nodes and connect two individual nodes u^{c_i} and v^{c_i} with an edge if the length of time two individuals spend in the same activity location is longer than δ hours (where $1 \leq \delta \leq 24$). As we know, the closer the individuals are to each other or the longer the individuals are in contact, the higher the risk of transmission. We call this graph $\mathcal{G}_A^{c_i}$ the activity-driven contact network. Fig. 4 illustrates an example of activity-driven contact network construction.

S4 In addition, to involve family-wise information in the contact network construction, we connect two individuals if they have the same household ID. We thus obtain a new contact network, i.e., an activity-family-driven contact network. Fig. 5 depicts an activity-family-driven contact network based on an activity-driven contact network from S3.

We apply our proposed activity-driven contact network construction and activity-family-driven contact network construction strategies on seven counties in the state of Virginia, i.e., Albemarle county, Charlotte county, Clarke county, Culpeper county, Henrico county, Highland county, and Prince Edward county. Table 1 presents basic county-level properties for the seven counties, and a specific one-month time period of COVID-19 transmission prediction for each county. Moreover, to better analyze the COVID-19 virus spreading, we consider building activity-driven contact and activity-family-driven contact networks with different connection densities. Specifically, inspired by ref. 32, we set the contact time δ (in hours) to be $\delta > 8$ h and $\delta > 12$ h. That is, we consider the lengths of time that two people come into contact with each other to be 8 and 12 h, which intuitively may be interpreted as the time spent together during normal business hours and time spent with families, respectively. Note that the actual time intervals for COVID-19 transmission may be lower, which will result in denser contact networks. However, the selected δ thresholds do not affect the generality of the proposed methodology and are chosen for illustrative purposes only. Hence, for each county, we can generate four types of contact networks, i.e., Activity-Driven Contact Network$_{8,\mathrm{h}}$ ($\mathcal{G}_A^{8h}$), Activity-Family-Driven Contact Network$_{8,\mathrm{h}}$ ($\mathcal{G}_{A\&F}^{8h}$), Activity-Driven Contact Network$_{12\,\mathrm{h}}$ ($\mathcal{G}_A^{12h}$), and Activity-Family-Driven Contact Network$_{12\,\mathrm{h}}$ ($\mathcal{G}_{A\&F}^{12h}$). Tables 2 and 3 show the summary statistics of the activity-driven contact network and activity-family-driven

Table 2. Summary of the synthetic social contact networks (with contact time > 8 h) in seven counties

	Network type	# Nodes	# Edges	# Triangles	Avg. deg.	Avg. density	Avg. $\mathcal{C}$	$\mathbf{k}_{\mathrm{avg}}/\mathbf{k}_{\mathrm{max}}$	Avg. household	Avg. age
Albemarle	$\mathcal{G}_A^{8h}$	27,988	307,082	284,678	21.944	7.841×10^{-4}	5.821×10^{-1}	16.539/210	2.036	34.101
	$\mathcal{G}_{A\&F}^{8h}$	55,698	362,949	307,074	13.033	2.339×10^{-4}	6.548×10^{-1}	9.901/210	2.266	31.326
Charlotte	$\mathcal{G}_A^{8h}$	3,484	23,571	20,758	13.531	3.885×10^{-3}	5.793×10^{-1}	10.481/86	2.377	35.524
	$\mathcal{G}_{A\&F}^{8h}$	7,812	35,704	28,292	9.141	1.170×10^{-3}	7.407×10^{-1}	7.249/86	2.795	35.316
Clarke	$\mathcal{G}_A^{8h}$	3,804	35,999	33,021	18.927	4.977×10^{-3}	5.743×10^{-1}	14.002/123	2.090	35.941
	$\mathcal{G}_{A\&F}^{8h}$	8,010	45,701	38,016	11.411	1.425×10^{-3}	6.900×10^{-1}	8.630/123	2.430	34.009
Culpeper	$\mathcal{G}_A^{8h}$	14,001	282,153	269,343	40.305	2.879×10^{-3}	5.908×10^{-1}	30.239/206	2.010	33.201
	$\mathcal{G}_{A\&F}^{8h}$	27,979	311,830	280,391	22.290	7.960×10^{-4}	6.647×10^{-1}	16.802/206	2.291	30.152
Hanover	$\mathcal{G}_A^{8h}$	28,819	459,960	435,851	31.921	1.107×10^{-3}	5.804×10^{-1}	23.923/221	1.923	34.460
	$\mathcal{G}_{A\&F}^{8h}$	58,986	515,855	462,866	17.491	2.965×10^{-4}	6.564×10^{-1}	13.169/221	2.148	31.641
Highland	$\mathcal{G}_A^{8h}$	1,306	9,948	8,843	15.234	1.167×10^{-2}	7.841×10^{-1}	14.442/76	2.240	41.546
	$\mathcal{G}_{A\&F}^{8h}$	1,585	11,240	9,571	14.183	8.953×10^{-3}	7.222×10^{-1}	12.732/76	2.513	40.903
Prince Edward	$\mathcal{G}_A^{8h}$	6,394	86,934	81,169	27.192	4.253×10^{-3}	5.619×10^{-1}	20.074/144	2.098	32.031
	$\mathcal{G}_{A\&F}^{8h}$	13,397	103,628	91,204	15.470	1.155×10^{-3}	6.916×10^{-1}	11.616/144	2.466	30.568

Table 3. Summary of the synthetic social contact networks (with contact time > 12 h) in seven counties

	Network type	# Nodes	# Edges	# Triangles	Avg. deg.	Avg. density	Avg. $\mathcal{C}$	$\mathbf{k}_{avg}/\mathbf{k}_{max}$	Avg. household	Avg. age
Albemarle	$\mathcal{G}_A^{12h}$	2,359	3,887	2,373	3.295	1.398×10^{-3}	3.473×10^{-1}	2.790/30	2.081	34.945
	$\mathcal{G}_{A\&F}^{12h}$	40,456	60,529	32,082	2.992	7.397×10^{-5}	7.448×10^{-1}	2.937/30	2.476	27.598
Charlotte	$\mathcal{G}_A^{12h}$	307	329	137	2.143	7.004×10^{-3}	3.119×10^{-1}	1.893/7	2.511	35.358
	$\mathcal{G}_{A\&F}^{12h}$	6,159	12,565	7,753	4.080	6.626×10^{-4}	8.702×10^{-1}	4.040/7	3.100	33.384
Clarke	$\mathcal{G}_A^{12h}$	330	385	181	2.333	7.092×10^{-3}	2.521×10^{-1}	2.009/12	2.030	38.558
	$\mathcal{G}_{A\&F}^{12h}$	5,980	10,209	5,831	3.414	5.711×10^{-4}	6.098×10^{-1}	3.373/12	2.730	30.687
Culpeper	$\mathcal{G}_A^{12h}$	1,146	2,144	1,389	3.742	3.268×10^{-3}	3.558×10^{-1}	3.108/16	2.105	33.825
	$\mathcal{G}_{A\&F}^{12h}$	20,591	32,355	17,682	3.142	1.526×10^{-4}	7.470×10^{-1}	3.073/16	2.537	26.195
Hanover	$\mathcal{G}_A^{12h}$	2,312	4,271	2,733	3.695	1.599×10^{-3}	3.724×10^{-1}	3.090/17	2.013	33.002
	$\mathcal{G}_{A\&F}^{12h}$	44,011	61,012	30,611	2.772	6.299×10^{-5}	7.470×10^{-1}	2.711/17	2.348	27.837
Highland	$\mathcal{G}_A^{12h}$	496	816	469	3.290	6.647×10^{-3}	4.802×10^{-1}	2.730/14	2.100	57.900
	$\mathcal{G}_{A\&F}^{12h}$	1,106	2,178	1,250	3.938	3.564×10^{-3}	6.533×10^{-1}	3.354/14	2.721	39.083
Prince Edward	$\mathcal{G}_A^{12h}$	641	1,539	1,119	4.802	7.503×10^{-3}	3.640×10^{-1}	3.970/22	2.129	32.062
	$\mathcal{G}_{A\&F}^{12h}$	10,065	18,426	10,938	3.661	3.638×10^{-4}	7.974×10^{-1}	3.573/22	2.771	28.002

contact network of each county with the contact time $\delta > 8$ h and $\delta > 12$ h, respectively. (Here Avg. $\mathcal{C}$ denotes the average clustering coefficient, and k_{avg} and k_{max} denote the average and maximal core number of each node, respectively). Figs. 6 and 7 show the visualization of activity-driven and activity-family-driven contact networks with the contact time $\delta > 12$ h of Albemarle county, Charlotte county, and Culpeper county, where the nodes are color coded to emphasize their degrees (the higher the degree of the node, the more red it is). Our data are available at https://github.com/SCPNAS/SC-SIR-Data.git. For the visualizations of other counties, see *SI Appendix*, section 8 and Figs. S7 and S8.

B. Results. We illustrate the utility of our proposed SC-SIR model in application to predict the number of COVID-19 infections among seven counties in Virginia. The data on the numbers of COVID-19 cases (i.e., used as ground truth) are obtained from NYTimes[†] via its COVID-19 data-gathering operations, and the original data have been collected by the Centers for Disease Control and Prevention.[‡]

The prediction periods of all seven counties are listed in Table 1. Prediction performance is measured by the Root Mean Square Error (RMSE). We also perform a one-sided two-sample t-test between the best result and the best performance achieved by the runner-up, where *, **, and *** are P-value < 0.1, 0.05, and 0.01 (i.e., denote significant, statistically significant, highly statistically, and significant results, respectively. We compare the prediction performance with the SIR model. The best results are bold. Furthermore, we also apply the SIR and SC-SIR models on $\mathcal{G}_A^{8h}$, $\mathcal{G}_{A\&F}^{8h}$, $\mathcal{G}_A^{12h}$, and $\mathcal{G}_{A\&F}^{12h}$, respectively. Note that, in our study, we calibrate the hyperparameters (including the transmission probabilities and recovery rate) for all models using grid-search cross-validation, and we use k-fold cross-validation (where $k = 5$).

The prediction performances on Activity-Driven Contact Network$_{8\,h}$ and Activity-Family-Driven Contact Network$_{8\,h}$ are reported in Table 2. We observe that

- Compared with the SIR model, the proposed SC-SIR always delivers better performance on all datasets. Especially, compared with the $\text{SIR}_{\mathcal{G}}^{A}$, $\text{SC-SIR}_{\mathcal{G}}^{AF}$ achieves maximum relative improvements of 94.556% on Culpeper county, 62.444% on Albemarle county, and 60.038% on Highland county. The results demonstrate the effectiveness of SC-SIR.
- $\text{SC-SIR}_{\mathcal{G}}^{AF}$ consistently outperforms $\text{SIR}_{\mathcal{G}}^{AF}$, and $\text{SC-SIR}_{\mathcal{G}}^{A}$ consistently outperforms $\text{SIR}_{\mathcal{G}}^{A}$ on all the datasets, indicating

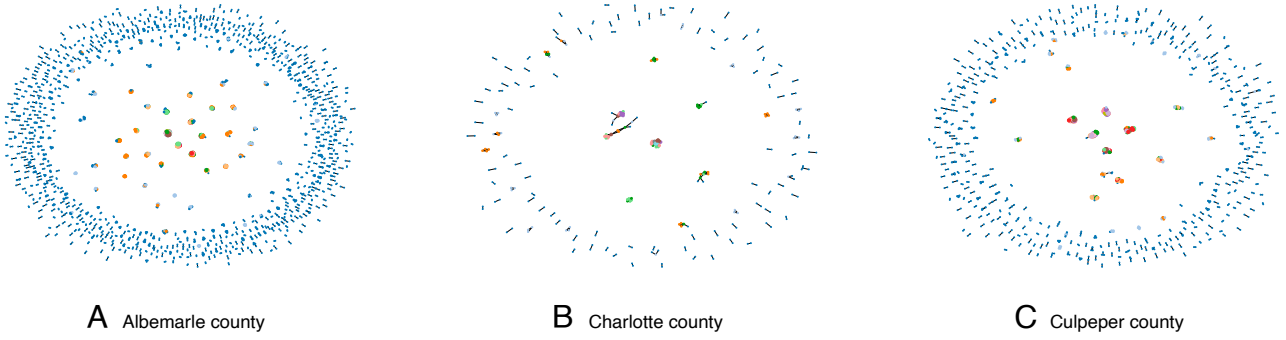

Fig. 6. Visualization of activity-driven contact networks with the contact time $\delta > 12$ h of (*A*) Albemarle county, (*B*) Charlotte county, and (*C*) Culpeper county. The higher the degree of the node, the more red it is.

[†]https://www.nytimes.com/interactive/2023/us/covid-cases.html.

[‡]https://covid.cdc.gov/covid-data-tracker/#datatracker-home.

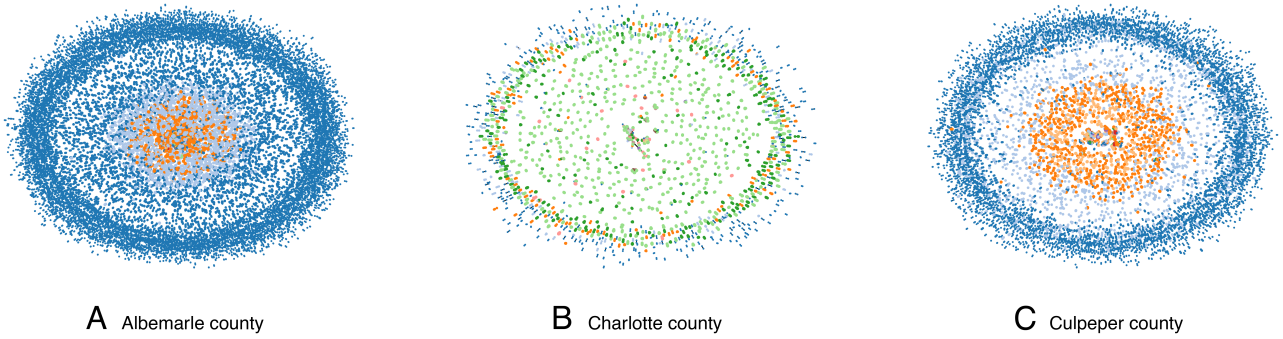

Fig. 7. Visualization of activity-family-driven contact networks with the contact time $\delta > 12$ h of (*A*) Albemarle county, (*B*) Charlotte county, and (*C*) Culpeper county. The higher the degree of the node, the more red it is.

the effectiveness of the higher-order interactions representation learning in SC-SIR.

- Comparing SC-SIR$_{\mathcal{G}}^{\mathrm{AF}}$ and SC-SIR$_{\mathcal{G}}^{\mathrm{A}}$, we observe that the SC-SIR model on the activity-family-driven contact network always shows better results than the SC-SIR model on the activity-driven contact network. This further confirms the importance of accounting for family factors in analysis of contagion dynamics such as COVID-19 virus.

Moreover, we found that there is a high correlation between higher-order structural features and the ratio of the infection rate (for both the node-wise infection rate and the higher-order infection rate) to the recovery rate. In particular, the higher the ratio of the number of triangles to the number of nodes, the higher ratio of the infection rate to the recovery rate is. A well-informed ratio of the infection rate to the recovery rate is important to establish the speed of spread and the effectiveness of interventions.

Table 5 presents the overall prediction performances, which are the averaged RMSE of our SC-SIR and SIR models on Activity-Driven Contact Network$_{12\,\mathrm{h}}$ and Activity-Family-Driven Contact Network$_{12\,\mathrm{h}}$ datasets. We can observe that i) the SC-SIR$_{\mathcal{G}}^{\mathrm{AF}}$ always significantly outperforms the SIR$_{\mathcal{G}}^{\mathrm{A}}$ model, ii) the average improvement of SC-SIR$_{\mathcal{G}}^{\mathrm{AF}}$ over SIR$_{\mathcal{G}}^{\mathrm{A}}$ on contact network with the contact time $\delta > 12$ h (i.e., 81.862%) is almost two times higher than that of SC-SIR$_{\mathcal{G}}^{\mathrm{AF}}$ over SIR$_{\mathcal{G}}^{\mathrm{A}}$ on contact network with the contact time $\delta > 8$ h (i.e., 44.464%), iii) the performances of SIR and SC-SIR models on activity-family-driven contact networks are consistently better than their performances on activity-driven contact networks. In addition, as shown in Tables 4 and 5, we find that the performances of our SC-SIR model on contact networks with contact time $\delta > 8$ h are generally better than the performances of our SC-SIR model on contact networks with contact time $\delta > 12$ h (except for Charlotte and Clarke counties), which indicates that the network density strongly impacts the performance of the SC-SIR model to predict epidemic spread. In light of the research Q2, these phenomena can be explained by the additional transmission channels which are introduced by the group interactions, and such new channels allow us to better capture heterogeneous properties of contagion dynamics.

Furthermore, we rank the improvements of our SC-SIR model on activity-family-driven contact networks with the contact times $\delta > 8$ h and $\delta > 12$ h. We observe that the gains of the SC-SIR model over the SIR model are related to the average density and average household size. For instance, the improvement of the SC-SIR over the SIR model in Hanover county is always lower than that of in other counties due to Hanover county exhibiting the lowest average density and the average number of people per household. Additionally, Tables 4 and 5 indicate that the SC-SIR model yields better prediction performance (i.e., smaller RMSE values) on Albemarle, Clarke, and Culpeper counties.

- Contact time > 8 h: Culpeper > Albemarle > Highland > Charlotte > Clarke > Prince Edward > Hanover;
- Contact time > 12 h: Highland > Charlotte > Clarke > Albemarle > Culpeper > Prince Edward > Hanover.

Figs. 8 and 9 show the prediction performance (i.e., the fraction of population infected at each timestamp) of the proposed SC-SIR model compared to the SIR model on activity-driven contact networks and activity-family-driven contact networks of seven counties, with the contact time $\delta > 8$ h and $\delta > 12$ h, respectively. Detailed experimental results on seven counties are presented in *SI Appendix*, section 9 and Tables S7–S12. Besides, we also provide the performance comparison in

Table 4. Performance comparison (RMSE) between SIR and SC-SIR on contact networks (where contact time > 8 h) in seven counties

	Activity-driven contact network$_{8\,\mathrm{h}}$		Activity-family-driven contact network$_{8\,\mathrm{h}}$		
County	SIR$_{\mathcal{G}}^{\mathrm{A}}$	SC-SIR$_{\mathcal{G}}^{\mathrm{A}}$	SIR$_{\mathcal{G}}^{\mathrm{AF}}$	SC-SIR$_{\mathcal{G}}^{\mathrm{AF}}$	Improvements %
Albemarle	1.595×10^{-3}	6.792×10^{-4}	7.903×10^{-4}	$^{***}\mathbf{6.407 \times 10^{-4}}$	62.444
Charlotte	4.677×10^{-3}	4.567×10^{-3}	3.615×10^{-3}	$^{**}\mathbf{3.215 \times 10^{-3}}$	31.259
Clarke	1.254×10^{-3}	1.079×10^{-3}	1.453×10^{-3}	$\mathbf{9.738 \times 10^{-4}}$	22.344
Culpeper	7.456×10^{-4}	6.136×10^{-4}	8.305×10^{-4}	$^{***}\mathbf{4.059 \times 10^{-4}}$	94.556
Hanover	1.941×10^{-3}	2.168×10^{-3}	1.576×10^{-3}	$^{**}\mathbf{1.570 \times 10^{-3}}$	19.114
Highland	5.809×10^{-3}	4.734×10^{-3}	2.542×10^{-3}	$^{**}\mathbf{2.322 \times 10^{-3}}$	60.028
Prince Edward	2.279×10^{-3}	1.926×10^{-3}	2.370×10^{-3}	$^{**}\mathbf{1.789 \times 10^{-3}}$	21.501

Table 5. Performance comparison (RMSE) between SIR and SC-SIR on contact networks (where contact time > 12 h) in seven counties

	Activity-driven contact network$_{12\,h}$		Activity-Family-driven contact network$_{12\,h}$		
County	$SIR^{A}_{\mathcal{G}}$	$SC\text{-}SIR^{A}_{\mathcal{G}}$	$SIR^{AF}_{\mathcal{G}}$	$SC\text{-}SIR^{AF}_{\mathcal{G}}$	Improvements %
Albemarle	1.706×10^{-3}	1.163×10^{-3}	9.096×10^{-4}	$^{**}\mathbf{6.623 \times 10^{-4}}$	61.178
Charlotte	8.963×10^{-3}	7.390×10^{-3}	4.693×10^{-3}	$^{***}\mathbf{2.884 \times 10^{-3}}$	67.823
Clarke	2.523×10^{-3}	2.380×10^{-3}	1.278×10^{-3}	$^{**}\mathbf{8.189 \times 10^{-4}}$	67.543
Culpeper	1.264×10^{-3}	1.642×10^{-3}	9.352×10^{-4}	$^{**}\mathbf{5.077 \times 10^{-4}}$	59.834
Hanover	2.962×10^{-3}	3.547×10^{-3}	2.511×10^{-3}	$^{**}\mathbf{1.609 \times 10^{-3}}$	45.679
Highland	1.435×10^{-2}	1.434×10^{-2}	5.021×10^{-3}	$^{***}\mathbf{4.567 \times 10^{-3}}$	218.258
Prince Edward	5.070×10^{-3}	4.072×10^{-3}	2.878×10^{-3}	$^{**}\mathbf{2.397 \times 10^{-3}}$	52.722

Pearson correlation coefficient in *SI Appendix*, section 5 and Table S4.

To better evaluate the effectiveness of our SC-SIR model, we compare the $SC\text{-}SIR_{\mathcal{G}}$ with the SIR ODE model on the activity-family-driven contact network with the contact time $\delta > 8$ h, i.e., SIR_{ODE}. Table 6 suggests that our $SC\text{-}SIR_{\mathcal{G}}$ always outperforms SIR_{ODE}. This observation indicates that modeling local graph structures and higher-order interactions is vital for understanding epidemic spreading. Furthermore, we present additional experiments on time-varying contact networks and apply the proposed SC-SIR model on contact networks with different sliding time windows (*SI Appendix*, sections 3 and 4). Note that while we currently focus on point forecasting, the proposed simplicial approach can be expanded to the probabilistic forecasting of the infection spread by using, for example, various forms of Kalman filters, Bayesian techniques, multi-model integration, and other ensemble based approaches (33–38). The resulting simplicial-based probabilistic forecasts can be evaluated using the weighted interval score (WIS) or other versions of proper scoring rules (39–43).

C. Mitigation Strategies. We now investigate the sensitivity of the proposed SC-SIR model to various mitigation strategies. In particular, we consider scenarios where more central nodes (individuals) are targeted to receive a vaccine, to quarantine or are persuaded to wear masks (44, 45). More specifically, we treat a mitigation strategy as a predefined fraction of nodes removed. For instance, if the $\tau\%$ nodes are selected in the decreasing order of their degree or betweenness, the resulting mitigation strategy is called a degree-based strategy or betweenness-based strategy, respectively. We consider two types of mitigation strategies, i.e., degree-based strategy and simplicial complex-based strategy. Similar to the degree-based strategy, for the proposed simplicial complex-based strategy, we first compute the amount of k-simplices (where $k = \{1, 2, \dots\}$) a node belongs to, and then rank all the nodes in descending order of the number of k-simplicies and select top $\tau\%$ influential nodes are considered as the most influential nodes under this strategy. We conduct experiments of the SC-SIR under three scenarios, i.e., without mitigation, degree-based mitigation, and simplicial complex-based mitigation. Specifically, for each mitigation, we remove

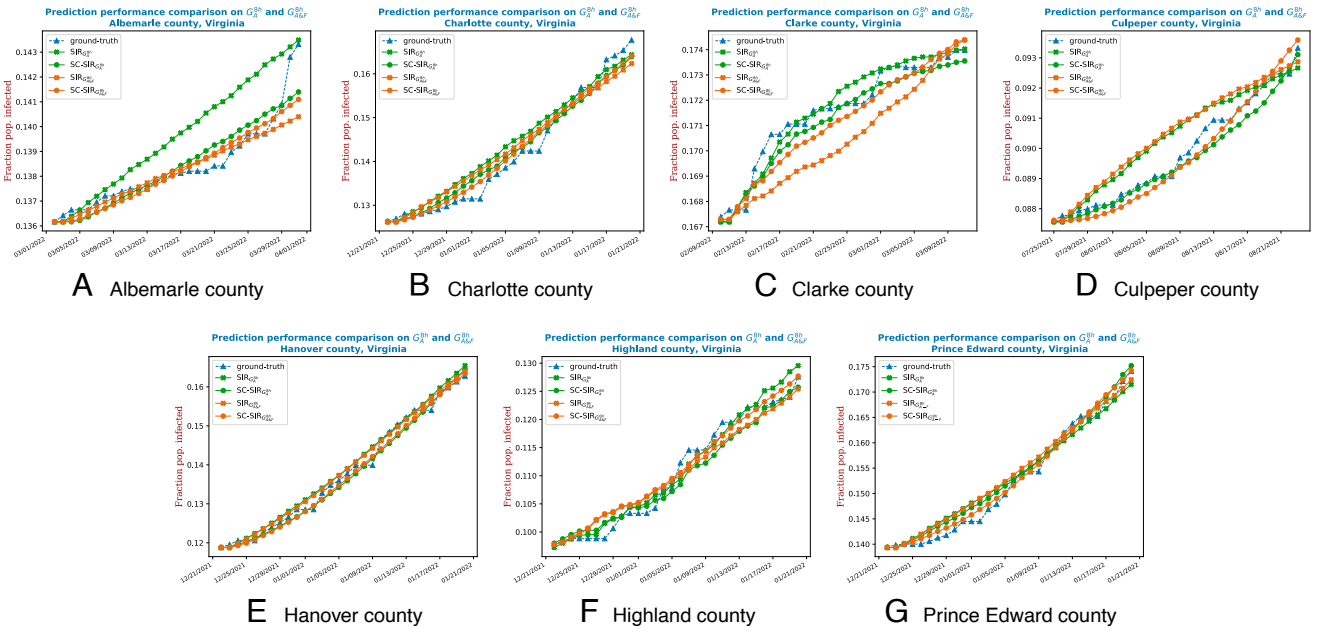

Fig. 8. Prediction performance comparison of SC-SIR and SIR models on activity-driven contact networks ($\mathcal{G}^{8h}_{A}$) and activity-family-driven contact networks ($\mathcal{G}^{8h}_{A\&F}$) with the contact time $\delta > 8$ h of seven counties, i.e., (*A*) Albemarle, (*B*) Charlotte, (*C*) Clarke, (*D*) Culpeper, (*E*) Hanover, (*F*) Highland, and (*G*) Prince Edward.

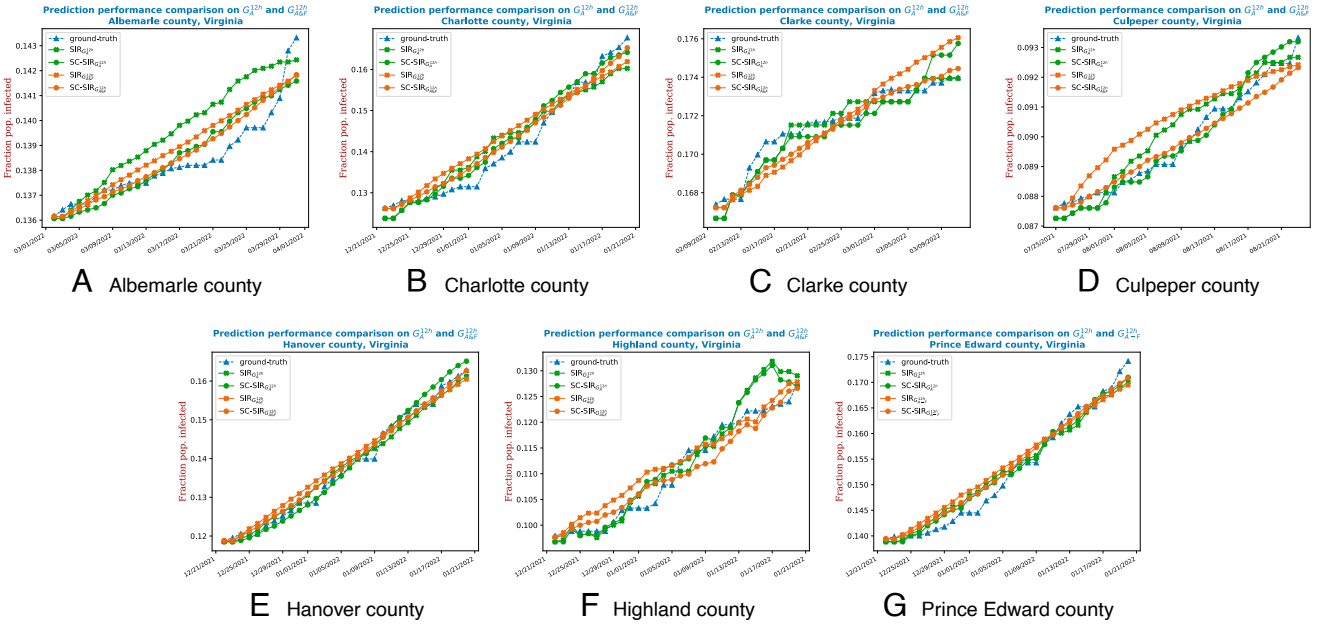

Fig. 9. Prediction performance comparison of SC-SIR and SIR models on activity-driven contact networks ($\mathcal{G}_A^{12h}$) and activity-family-driven contact networks ($\mathcal{G}_{A\&F}^{12h}$) with the contact time $\delta > 12$ h of seven counties, i.e., (*A*) Albemarle, (*B*) Charlotte, (*C*) Clarke, (*D*) Culpeper, (*E*) Hanover, (*F*) Highland, and (*G*) Prince Edward.

the nodes with top $\tau\%$ node degree scores or top $\tau\%$ amount of k-simplices (here, we set k to be 2). Table 7 shows the performance comparisons of our SC-SIR and SIR models under three scenarios (i.e., without mitigation, degree-based mitigation, and simplicial complex-based mitigation) on all seven counties. As Table 7 suggests, compared with the SIR, the SC-SIR model can get better performances on all scenarios. Moreover, through comparing the performances of $\text{SIR}_{\mathcal{G}}^{\text{d}}$ and $\text{SIR}_{\mathcal{G}}^{\text{sc}}$ (i.e., the SC-SIR models under degree-based and simplicial complex-based mitigation strategies respectively), the results demonstrate that $\text{SIR}_{\mathcal{G}}^{\text{sc}}$ achieves better performance across all counties except for Albemarle county. In light of the research Q4, these findings suggest that, first, the higher-order interactions play an important role in the COVID-19 spread and, second, the simplicial complex-based approach for disease control may be a competitive alternative for existing mitigation strategies (44).

D. From SC-SIR to SC-SEIR. Both SIR and Susceptible-Exposed-Infected-Recovered (SEIR) models are widely used in epidemiology to describe the spread of infectious diseases. Different from the SIR model, the SEIR model adds an additional compartment, exposed, which represents individuals who have been infected but are not yet infectious. Specifically, individuals become exposed after exposure, and then become infectious after a latent period during which the virus replicates in their bodies. To study the effectiveness of our proposed simplicial complex-based framework, in this section, we extend our SC-SIR to a simplicial complex-based SEIR (i.e., SC-SEIR$_{\mathcal{G}}$) model for the long-term prediction. Moreover, instead of working on datasets shown in Tables 2 and 3, we apply SC-SEIR$_{\mathcal{G}}$ and SEIR$_{\mathcal{G}}$ models on *specific* activity-family-driven contact networks. From the activity location assignment file, there are four different types of activity locations, and then we build the activity-family-driven contact network for each type of activity location separately by using the contact network construction scheme (following the section A and we set the contact time $\delta > 8$ h). Hence, we obtain the i) shopping-family-driven contact network ($\mathcal{G}_{Sho\&F}^{8h}$), ii) school-family-driven contact network ($\mathcal{G}_{Sch\&F}^{8h}$), iii) college-family-driven contact network ($\mathcal{G}_{C\&F}^{8h}$), and iv) religion-family-driven contact network ($\mathcal{G}_{R\&F}^{8h}$) for four types of locations respectively. We then apply both SEIR$_{\mathcal{G}}$ and SC-SEIR$_{\mathcal{G}}$ models to all the above specific contact networks in Albemarle county, Charlotte county, and Hanover county. Tables 8–10 display the overall prediction performances of our SC-SEIR$_{\mathcal{G}}$ and SEIR$_{\mathcal{G}}$ models on specific contact networks of three counties from 01/22/2020 to 07/19/2020 (totaling 180 d). We observe that the proposed SC-SEIR$_{\mathcal{G}}$ model achieves better performance across all three datasets. For instance, our SC-SEIR$_{\mathcal{G}}$ on $\mathcal{G}_{R\&F}^{8h}$ delivers relative gains of 25.0%, 27.0%, and 8.62% on Albemarle county, Charlotte county, and Hanover county respectively. To sum up, our results indicate that information on higher-order interactions is an important complementary asset leading to improved predictive performance.

E. From SC-SIR to SC-SEIAR. In the real world, the spread of COVID-19 includes not only symptomatic individuals, but also those who do not show any symptoms (i.e., asymptomatic), Thus, in this study, we also incorporate asymptomatic infections into

Table 6. The estimation comparison on seven counties between SIR_{ODE} and SC-SIR$_{\mathcal{G}}$

Model	Albemarle	Charlotte	Clarke	Culpeper	Hanover	Highland	Prince Edward
SC-SIR$_{\mathcal{G}}$	$\mathbf{6.407 \times 10^{-4}}$	$\mathbf{2.884 \times 10^{-3}}$	$\mathbf{8.189 \times 10^{-4}}$	$\mathbf{4.059 \times 10^{-4}}$	$\mathbf{1.570 \times 10^{-3}}$	$\mathbf{2.322 \times 10^{-3}}$	$\mathbf{1.789 \times 10^{-3}}$
SIR_{ODE}	7.208×10^{-4}	9.679×10^{-2}	9.371×10^{-4}	4.137×10^{-4}	2.255×10^{-3}	2.330×10^{-3}	3.024×10^{-3}

Table 7. Mitigation strategies comparison on seven counties

	Without mitigation		Degree-based mitigation (10%)		Simplicial complex-based mitigation (10%)	
County	$SIR_{\mathcal{G}}$	$SC\text{-}SIR_{\mathcal{G}}$	$SIR^{d}_{\mathcal{G}}$	$SC\text{-}SIR^{d}_{\mathcal{G}}$	$SIR^{sc}_{\mathcal{G}}$	$SC\text{-}SIR^{sc}_{\mathcal{G}}$
Albemarle	2.630×10^{-3}	2.239×10^{-3}	1.273×10^{-3}	1.161×10^{-3}	1.666×10^{-3}	1.416×10^{-3}
Charlotte	3.545×10^{-3}	3.418×10^{-3}	3.253×10^{-3}	3.230×10^{-3}	3.244×10^{-3}	3.166×10^{-3}
Clarke	3.232×10^{-3}	3.034×10^{-3}	3.145×10^{-3}	2.834×10^{-3}	2.562×10^{-3}	2.516×10^{-3}
Culpeper	3.537×10^{-3}	2.809×10^{-3}	2.302×10^{-3}	2.143×10^{-3}	2.044×10^{-3}	1.823×10^{-3}
Hanover	2.147×10^{-3}	2.062×10^{-3}	1.864×10^{-3}	1.738×10^{-3}	1.667×10^{-3}	1.610×10^{-3}
Highland	1.003×10^{-2}	8.170×10^{-3}	9.123×10^{-3}	7.525×10^{-3}	6.734×10^{-3}	6.541×10^{-3}
Prince Edward	3.672×10^{-3}	3.618×10^{-3}	2.249×10^{-3}	2.054×10^{-3}	1.940×10^{-3}	1.872×10^{-3}

Table 8. Overall prediction performance of $SEIR_{\mathcal{G}}$ and $SC\text{-}SEIR_{\mathcal{G}}$ on Albemarle county

Model	$\mathcal{G}^{8h}_{Sho\&F}$	$\mathcal{G}^{8h}_{Sch\&F}$	$\mathcal{G}^{8h}_{C\&F}$	$\mathcal{G}^{8h}_{R\&F}$
$SC\text{-}SEIR_{\mathcal{G}}$	**0.232**	**0.678**	**0.072**	**0.168**
$SEIR_{\mathcal{G}}$	0.253	0.712	0.086	0.210

Table 9. Overall prediction performance of $SEIR_{\mathcal{G}}$ and $SC\text{-}SEIR_{\mathcal{G}}$ on Charlotte county

Model	$\mathcal{G}^{8h}_{Sho\&F}$	$\mathcal{G}^{8h}_{Sch\&F}$	$\mathcal{G}^{8h}_{C\&F}$	$\mathcal{G}^{8h}_{R\&F}$
$SC\text{-}SEIR_{\mathcal{G}}$	**0.124**	**0.534**	**0.068**	**0.174**
$SEIR_{\mathcal{G}}$	0.167	0.593	0.069	0.221

Table 10. Overall prediction performance of $SEIR_{\mathcal{G}}$ and $SC\text{-}SEIR_{\mathcal{G}}$ on Hanover county

Model	$\mathcal{G}^{8h}_{Sho\&F}$	$\mathcal{G}^{8h}_{Sch\&F}$	$\mathcal{G}^{8h}_{C\&F}$	$\mathcal{G}^{8h}_{R\&F}$
$SC\text{-}SEIR_{\mathcal{G}}$	**0.107**	**0.102**	**0.085**	**0.106**
$SEIR_{\mathcal{G}}$	0.191	0.115	0.097	0.116

Table 11. Overall prediction performance of $SEIAR_{\mathcal{G}}$ and $SC\text{-}SEIAR_{\mathcal{G}}$ on Albemarle county

Model	$\mathcal{G}^{8h}_{Sho\&F}$	$\mathcal{G}^{8h}_{Sch\&F}$	$\mathcal{G}^{8h}_{C\&F}$	$\mathcal{G}^{8h}_{R\&F}$
$SC\text{-}SEIAR_{\mathcal{G}}$	**0.213**	**0.576**	**0.072**	**0.136**
$SEIAR_{\mathcal{G}}$	0.251	0.649	0.074	0.198

$SEIR_{\mathcal{G}}$ and introduce a model, namely $SEIAR_{\mathcal{G}}$ (Susceptible–Exposed–Infectious–Asymptomatic–Recovered) model. Specifically, different from the $SEIR_{\mathcal{G}}$ model, if susceptible persons contact with asymptomatic and symptomatic infected people, their status will change to exposed. Specifically, after the exposed period, the probability of an exposed person entering into symptomatic infected status is β and the probability of being exposed into asymptomatic infected is β' (where $\beta + \beta' = 1$). Similar to the $SC\text{-}SEIR_{\mathcal{G}}$ model (see subsection D), in this subsection, we develop a simplicial complex-based SEIAR (i.e., $SC\text{-}SEIAR_{\mathcal{G}}$) model for the COVID-19 transmission prediction on Albemarle, Charlotte, and Hanover counties. In Tables 11–13, we show the prediction performance of our $SC\text{-}SEIAR_{\mathcal{G}}$ and $SEIAR_{\mathcal{G}}$ on three counties. Clearly, the proposed $SC\text{-}SEIAR_{\mathcal{G}}$ always performs better than $SEIAR_{\mathcal{G}}$ on all datasets cross different types of contact networks. Moreover, we observe that $SC\text{-}SEIAR_{\mathcal{G}}$ achieves a better performance compared with the $SC\text{-}SEIR_{\mathcal{G}}$. For example, on Charlotte county especially, $SC\text{-}SEIAR_{\mathcal{G}}$ yields more than 3.57% relative improvements to the $SC\text{-}SEIR_{\mathcal{G}}$, hence demonstrating the effectiveness of our method for the prediction of COVID-19 cases.

Table 12. Overall prediction performance of $SEIAR_{\mathcal{G}}$ and $SC\text{-}SEIAR_{\mathcal{G}}$ on Charlotte county

Model	$\mathcal{G}^{8h}_{Sho\&F}$	$\mathcal{G}^{8h}_{Sch\&F}$	$\mathcal{G}^{8h}_{C\&F}$	$\mathcal{G}^{8h}_{R\&F}$
$SC\text{-}SEIR_{\mathcal{G}}$	**0.113**	**0.497**	**0.056**	**0.168**
$SEIAR_{\mathcal{G}}$	0.187	0.539	0.077	0.200

Table 13. Overall prediction performance of $SEIAR_{\mathcal{G}}$ and $SC\text{-}SEIAR_{\mathcal{G}}$ on Hanover county

Model	$\mathcal{G}^{8h}_{Sho\&F}$	$\mathcal{G}^{8h}_{Sch\&F}$	$\mathcal{G}^{8h}_{C\&F}$	$\mathcal{G}^{8h}_{R\&F}$
$SC\text{-}SEIAR_{\mathcal{G}}$	**0.102**	**0.098**	**0.079**	**0.129**
$SEIAR_{\mathcal{G}}$	0.185	0.108	0.096	0.126

4. Conclusion and Discussion

The goal of this paper is to glean a better and more systematic understanding of the potential role that various higher-order social group interactions may play in contagion propagation—a fundamental problem that has recently received a surge of interest in a broad range of disciplines, from biosurveillance to computer science. To design a more holistic approach to this open problem and to learn the additional transmission routes, we have capitalized on the emerging concepts of simplicial models. Our results have confirmed the intuitive premise that *explicitly* accounting for group (or more formally, polyadic) interactions play an important role in both tracking the contagion spread and developing more efficient mitigation strategies. In particular, integrating simplicial complexes which describe such higher-order social interactions has led to a reduction of RMSE from 0.09% to 25.75% over seven counties in the state of Virginia. This phenomenon has been found to manifest across all considered types of mechanistic models, and, as it could be expected, the impact of group interactions increases with the joint time the group members spend together. Not surprisingly then, family-driven activities have demonstrated the highest impact on the contagion dynamics.

While the obtained results offer an important glimpse into the hidden mechanisms behind contagion propagation on social networks, this study is yet just one of the first steps toward a better understanding of the key driving factors of spread dynamics. In particular, little is known about uncertainty propagation associated with simplicial models. One of the promising approaches here is to use arbitrary polynomial chaos expansion on simplicial complex-based mechanistic models, thereby, extending the recent results of ref. 46. Furthermore, it is important to investigate the critical time the group can spend together in order to exhibit a substantial impact on the

spread dynamics. Such impact in turn is closely related to the imposed mitigation measures such as curfews, quarantines and the associated mobility patterns (47–50). Finally, the higher-order group interactions are to be systematically integrated into the construction of synthetic social contact networks (51–57) and, more generally, be accounted in digital twins of population behavior. Needless to say, these implications and open questions are valid way beyond biosurveillance and apply to a wide range of problems, from information propagation and signal processing (58, 59) to financial contagion and fraud detection (60, 61) to resiliency of critical infrastructures (62–64).

Finally, we would like to emphasize that all the discussed models, from the independent cascade approaches to simplicial ones, are simply mathematical abstractions for describing the virus spread (65). Our premise is that the simplicial abstraction and its systematic treatment under various realistic scenarios may open a path for a better understanding of various yet largely unexplained phenomena about the latent relationships between the level of dose and risk of infection.

Data, Materials, and Software Availability. All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. Y.C. has been supported by the NASA AIST grant 21-AIST21_2-0059, NSF Grants DMS-2335846/2335847, TIP-2333703. Y.R.G. has been supported by the ONR award N00014-21-1-2530 and NASA AIST grants 21-AIST21_2-0020 and 21-AIST21_2-0059. M.V.M has been supported in part by the following grants: University of Virginia Strategic Investment Fund (Award No. SIF160), NSF Grants CCF-1918656 (Expeditions), OAC-1916805 (CINES), IIS-1955797, VDH Grant PV-BII VDH COVID-19 Modeling Program VDH-21-501-0135, DTRA subcontract/ARA S-D00189-15-TO-01-UVA. H.V.P has been supported by NSF Grant ECCS-2039716. Part of this material is also based upon work supported by (while Y.R.G. was serving at) the NSF. The views expressed in the article do not necessarily represent the views of NSF.

Author affiliations: [a]Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122; [b]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080; [c]Division of Mathematical Sciences, NSF, Alexandria, VA 22314; [d]Department of Computer Science, University of Virginia; [e]Biocomplexity Institute and Initiative, University of Virginia, Charlottesville, VA 22904; and [f]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ 08544

1. A. R. Schaub, B. P. Horn, G. Lippner, A. Jadbabaie, Random walks on simplicial complexes and the normalized Hodge 1-Laplacian. *SIAM Rev.* **62**, 353–391 (2020).
2. L. Torres, A. S. Blevins, D. Bassett, T. Eliassi-Rad, The why, how, and when of representations for complex systems. *SIAM Rev.* **63**, 435–485 (2021).
3. S. Majhi, M. Perc, D. Ghosh, Dynamics on higher-order networks: A review. *J. R. Soc. Interface* **19**, 20220043 (2022).
4. T. Gibbs, S. A. Levin, J. M. Levine, Coexistence in diverse communities with higher-order interactions. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2205063119 (2022).
5. R. Herzog *et al*., Genuine high-order interactions in brain networks and neurodegeneration. *Neurobiol. Dis.* **175**, 105918 (2022).
6. M. S. Pour, C. Nader, K. Friday, E. Bou-Harb, A comprehensive survey of recent internet measurement techniques for cyber security. *Comput. Sec.* **128**, 103123 (2023).
7. Z. Yan, T. Ma, L. Gao, Z. Tang, C. Chen, "Link prediction with persistent homology: An interactive view" in *Proceedings of the International Conference on Machine Learning* (2021), pp. 11659–11669.
8. I. Iacopini, G. Petri, A. Barrat, V. Latora, Simplicial models of social contagion. *Nat. Commun.* **10**, 2485 (2019).
9. J. T. Matamalas, S. Gómez, A. Arenas, Abrupt phase transition of epidemic spreading in simplicial complexes. *Phys. Rev. Res.* **2**, 012049 (2020).
10. Z. Li *et al*., Contagion in simplicial complexes. *Chaos, Solitons Fractals* **152**, 111307 (2021).
11. Y. Chen, Y. R. Gel, H. V. Poor, "BScNets: Block simplicial complex neural networks" in *Proceedings of the AAAI Conference on Artificial Intelligence* (2022), vol. 36, pp. 6333–6341.
12. G. Burgio, A. Arenas, S. Gómez, J. T. Matamalas, Network clique cover approximation to analyze complex contagions through group interactions. *Commun. Phys.* **4**, 111 (2021).
13. F. Battiston *et al*., Networks beyond pairwise interactions: Structure and dynamics. *Phys. Rep.* **874**, 1–92 (2020).
14. N. W. Landry, J. G. Restrepo, The effect of heterogeneity on hypergraph contagion models. *Chaos* **30**, 103117 (2020).
15. S. Sinha, S. Bhattacharya, S. Roy, Impact of second-order network motif on online social networks. *J. Supercomput.* **78**, 5450–5478 (2022).
16. M. Ritchie, L. Berthouze, T. House, I. Z. Kiss, Higher-order structure and epidemic dynamics in clustered networks. *J. Theor. Biol.* **348**, 21–32 (2014).
17. A. K. Dey, Y. R. Gel, H. V. Poor, What network motifs tell us about resilience and reliability of complex networks. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 19368–19373 (2019).
18. Q. F. Lotito, F. Musciotto, A. Montresor, F. Battiston, Higher-order motif analysis in hypergraphs. *Commun. Phys.* **5**, 79 (2022).
19. D. Wang, Y. Zhao, J. Luo, H. Leng, Simplicial sirs epidemic models with nonlinear incidence rates. *Chaos* **31**, 053112 (2021).
20. P. S. Dodds, D. J. Watts, A generalized model of social and biological contagion. *J. Theor. Biol.* **232**, 587–604 (2005).
21. D. Guilbeault, J. Becker, D. Centola, "Complex contagions: A decade in review" in *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks* (2018), pp. 3–25.
22. W. Van Damme, R. Dahake, R. Van de Pas, G. Vanham, Y. Assefa, COVID-19: Does the infectious inoculum dose–response relationship contribute to understanding heterogeneity in disease severity and transmission dynamics? *Med. Hypothes.* **146**, 110431 (2021).
23. L. M. Brosseau *et al*., Severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) dose, infection, and disease outcomes for coronavirus disease 2019 (COVID-19): A review. *Clin. Infect. Dis.* **75**, e1195–e1201 (2022).
24. R. Ke, C. Zitzmann, D. D. Ho, R. M. Ribeiro, A. S. Perelson, In vivo kinetics of SARS-COV-2 infection and its relationship with a person's infectiousness. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2111477118 (2021).
25. K. Koelle *et al*., Masks do no more than prevent transmission: Theory and data undermine the variolation hypothesis. medRxiv [Preprint] (2022). https://doi.org/10.1101/2022.06.28.22277028.
26. A. Adiga *et al*., Strategies to mitigate COVID-19 resurgence assuming immunity waning: A study for Karnataka, India. medRxiv [Preprint] (2021). https://doi.org/10.1101/2021.05.26.21257836.
27. K. B. Law *et al*., Tracking the early depleting transmission dynamics of COVID-19 with a time-varying SIR model. *Sci. Rep.* **10**, 21721 (2020).
28. Z. Liao, P. Lan, Z. Liao, Y. Zhang, S. Liu, TW-SIR: Time-window based SIR for COVID-19 forecasts. *Sci. Rep.* **10**, 22454 (2020).
29. M. Kiamari *et al*., "COVID-19 risk estimation using a time-varying SIR-model" in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Modeling and Understanding the Spread of COVID-19* (2020), pp. 36–42.
30. Z. Peng *et al*., Estimating unreported COVID-19 cases with a time-varying SIR regression model. *Int. J. Environ. Res. Public Health* **18**, 1090 (2021).
31. S. Vecherin *et al*., Assessment of the COVID-19 infection risk at a workplace through stochastic microexposure modeling. *J. Exp. Sci. Environ. Epidemiol.* **32**, 712–719 (2022).
32. S. Pei, S. Kandula, J. Shaman, Differential effects of intervention timing on COVID-19 spread in the United States. *Sci. Adv.* **6**, eabd6370 (2020).
33. A. Adiga *et al*., "All models are useful: Bayesian ensembling for robust high resolution COVID-19 forecasting" in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 2505–2513.
34. L. L. Ramírez-Ramírez, Y. R. Gel, M. Thompson, E. de Villa, M. McPherson, A new surveillance and spatio-temporal visualization tool SIMID: SIMulation of infectious diseases using random networks and GIS. *Comput. Methods Programs Biomed.* **110**, 455–470 (2013).
35. N. G. Reich *et al*., A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 3146–3154 (2019).
36. A. Adiga *et al*., "AI techniques for forecasting epidemic dynamics: Theory and practice" in *Artificial Intelligence in COVID-19*, N. Lidströmer, Y. C. Eldar, Eds. (Springer, 2022), pp. 193–228.
37. J. Botz *et al*., Modeling approaches for early warning and monitoring of pandemic situations as well as decision support. *Front. Public Health* **10**, 994949 (2022).
38. M. A. Achterberg *et al*., Comparing the accuracy of several network-based COVID-19 prediction algorithms. *Int. J. Forecast.* **38**, 489–504 (2022).
39. T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
40. A. Adiga *et al*., Evaluating the impact of international airline suspensions on the early global spread of COVID-19. medRxiv [Preprint] (2020). https://doi.org/10.1101/2020.02.20.20025882.
41. J. Bracher, E. L. Ray, T. Gneiting, N. G. Reich, Evaluating epidemic forecasts in an interval format. *PLoS Comput. Biol.* **17**, e1008618 (2021).
42. A. Adiga *et al*., "Enhancing COVID-19 ensemble forecasting model performance using auxiliary data sources" in *2022 IEEE International Conference on Big Data* (2022), pp. 1594–1603.
43. E. Y. Cramer *et al*., Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the united states. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2113561119 (2022).
44. J. Chen *et al*., Prioritizing allocation of covid-19 vaccines based on social contacts increases vaccination effectiveness. medRxiv [Preprint] (2021). https://doi.org/10.1101/2021.02.04.21251012.
45. R. P. Curiel, H. G. Ramírez, Vaccination strategies against COVID-19 and the diffusion of anti-vaccination views. *Sci. Rep.* **11**, 1–13 (2021).
46. S. Ubaru, L. Horesh, G. Cohen, Dynamic graph and polynomial chaos based models for contact tracing data analysis and optimal testing prescription. *J. Biomed. Inf.* **122**, 103901 (2021).
47. S. Chang *et al*., "Supporting COVID-19 policy response with large-scale mobility-based modeling" in *Proceedings of the SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 2632–2642.
48. A. Adiga *et al*., Impact of weeknight and weekend curfews using mobility data: A case study of Bengaluru Urban. medRxiv [Preprint] (2022). https://doi.org/10.1101/2022.01.26.22269903.
49. S. Venkatramanan *et al*., Forecasting influenza activity using machine-learned mobility map. *Nat. Commun.* **12**, 726 (2021).

50. A. K. Ligo *et al*., Relationship among state reopening policies, health outcomes and economic recovery through first wave of the COVID-19 pandemic in the US. *PLoS ONE* **16**, e0260015 (2021).
51. C. L. Barrett *et al*., "Generation and analysis of large synthetic social contact networks" in *Proceedings of the 2009 IEEE Winter Simulation Conference* (2009), pp. 1003–1014.
52. K. R. Bisset, X. Feng, M. Marathe, S. Yardi, "Modeling interaction between individuals, social networks and public policy to support public health epidemiology" in *Proceedings of the 2009 IEEE Winter Simulation Conference* (2009), pp. 2020–2031.
53. N. Geard, J. M. McCaw, A. Dorin, K. B. Korb, J. McVernon, Synthetic population dynamics: A model of household demography. *J. Artif. Soc. Soc. Simul.* **16**, 8 (2013).
54. S. Bates, V. Leonenko, J. Rineer, G. Bobashev, Using synthetic populations to understand geospatial patterns in opioid related overdose and predicted opioid misuse. *Comput. Math. Org. Theory* **25**, 36–47 (2019).
55. L. Wang, J. Chen, M. Marathe, TDEFSI: Theory-guided deep learning-based epidemic forecasting with synthetic information. *ACM Trans. Spat. Algorithms Syst.* **6**, 1–39 (2020).
56. G. Papyshev, M. Yarime, Exploring city digital twins as policy tools: A task-based approach to generating synthetic data on urban mobility. *Data Policy* **3**, e16 (2021).
57. Ç. Tozluoğlu *et al*., A synthetic population of Sweden: Datasets of agents, households, and activity-travel patterns. *Data Brief* **48**, 109209 (2023).
58. I. Iacopini, G. Petri, A. Baronchelli, A. Barrat, Group interactions modulate critical mass dynamics in social convention. *Commun. Phys.* **5**, 64 (2022).
59. M. T. Schaub *et al*., "Signal processing on simplicial complexes" in *Higher-Order Systems*, F. Battiston, G. Petri, Eds. (Springer, 2022), pp. 301–328.
60. M. Y. Korniyenko, M. Patnam, R. M. del Rio-Chanon, M. A. Porter, *Evolution of the Global Financial Network and Contagion: A New Approach* (International Monetary Fund, 2018).
61. C. G. Akcora, Y. Li, Y. R. Gel, M. Kantarcioglu, "Bitcoinheist: Topological data analysis for ransomware detection on the bitcoin blockchain" in *Proceedings of the International Joint Conference on Artificial Intelligence* (2020).
62. R. Meyur *et al*., "Creating realistic power distribution networks using interdependent road infrastructure" in *2020 IEEE International Conference on Big Data* (2020), pp. 1226–1235.
63. A. K. Dey, S. J. Young, Y. R. Gel, From Delaunay triangulation to topological data analysis: Generation of more realistic synthetic power grid networks. *J. R. Stat. Soc., Ser. A: Stat. Soc.* **186**, 335–354 (2023).
64. R. Meyur *et al*., Ensembles of realistic power distribution networks. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2205772119 (2022).
65. D. J. Rosenkrantz *et al*., Fundamental limitations on efficiently forecasting certain epidemic measures in network models. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2109228119 (2022).