The Interplay of Clustering and Evolution in the Emergence of Epidemics on Networks

Mansi Sood[γ], Rashad Eletreby[†], Swarun Kumar[γ], Chai Wah Wu[+], Osman Yağan[γ]
[γ]Department of Electrical and Computer Engineering and CyLab,
Carnegie Mellon University, Pittsburgh, PA 15213 USA
[†]Rocket Travel, Inc, Chicago, IL 60661 USA
[+]Thomas J. Watson Research Center, IBM, Yorktown Heights, NY 10598 USA
msood@andrew.cmu.edu, eletreby.rashad@gmail.com, swarun@cmu.edu, cwwu@us.ibm.com, oyagan@andrew.cmu.edu

ICC 2023 - IEEE International Conference on Communications | 978-1-5386-7462-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICC45041.2023.10279814

***Abstract*—We are living amidst a pandemic caused by a ravaging coronavirus and an accompanying pandemic of misinformation that has strained our economy and socio-political institutions. A key scientific goal is to examine mechanisms that lead to the widespread propagation of contagions, e.g., misinformation and pathogens, and identify risk factors that can trigger widespread outbreaks. A common phenomenon underlying the spread of disease and misinformation epidemics is the *evolution* of the contagion as it propagates, leading to the emergence of different *strains*, e.g., through genetic mutations in pathogens and alterations in the information content. Recent studies have revealed that models that do not account for heterogeneity in transmission risks associated with different strains of the circulating contagion can lead to inaccurate predictions. However, existing results on *multi-strain* spreading assume that the network has a vanishingly small *clustering* coefficient, whereas, clustering is widely known to be a fundamental property of real-world social networks.**

In this work, we investigate spreading processes that entail evolutionary adaptations on random graphs with tunable clustering and arbitrary degree distributions. We derive a mathematical framework that predicts the epidemic threshold and the probability of emergence as functions of the characteristics of the spreading object, the evolutionary pathways of the pathogen/misinformation, and the structure of the underlying network as given by the *joint* degree distribution of single-edges and triangles. To the best of our knowledge, our work is the first to jointly characterize the impact of clustering and evolution on the emergence of epidemic outbreaks. We supplement our theoretical finding with numerical simulations and case studies, shedding light on how clustering can offer pathways for mutation, thereby altering the course of the epidemic.

***Index Terms*—Spreading Processes, Clustering, Evolution, Agent-based models, Social Networks, Epidemics,**

I. Introduction

A. Background and Related Work

The recent outbreak of COVID-19 triggered by the novel coronavirus SARS-CoV-2 led to a widespread strain on public health and the economy. The highly transmissible and rapidly *evolving* [1] nature of the SARS-CoV-2 coronavirus and the absence of pharmacological interventions in the early stages of the outbreak led to lock-downs and social distancing measures to combat the spread. The widespread use of social media platforms to trigger misinformation pertaining to COVID-19 further impeded efforts to mitigate its spread [2]. Spreading processes have emerged as a key analytical and computational tool for understanding the mechanisms underlying the spread of contagions in different contexts, e.g., the spread of infectious diseases and misinformation.

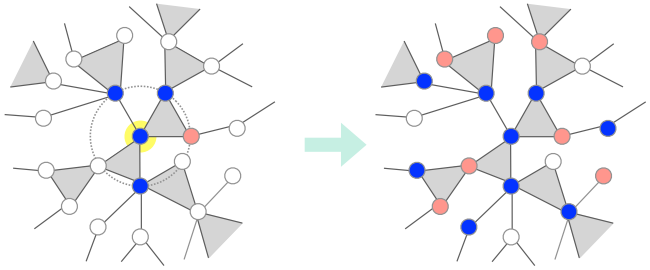

Fig. 1: *The contact network comprises a clustered network where the number of single-edges and triangles attached to each node is separately specified through the joint degree distribution. We consider the propagation of two strains of a contagion indicated in blue and red. The spreading proceeds as follows: An arbitrarily chosen seed node acquires strain-1. The seed node independently infects its susceptible neighbors with a probability T_1. After infection, the contagion mutates to strain-2 within the hosts with probability μ_{12}. The process continues recursively and terminates when no further infections are possible.*

Akin to different *strains* of a pathogen arising through evolutionary adaptations or mutations, different *versions* of the information are created as the content is altered on social media platforms [3], [4]. The resulting *variants/strains* of the information may have varying propensities to be circulated in the social network. A key scientific goal across the above spreading phenomena is to examine the underlying mechanisms that lead to the widespread propagation of contagions, e.g., misinformation and pathogens, and identify risk factors that can trigger widespread outbreaks. From the standpoint of spreading processes, this necessitates the development of a rich class of agent-based models that can simultaneously account for realistic patterns of interaction in the population and evolution in the contagion.

Most network-based epidemiological models do not consider the role of evolutionary adaptations in the spread of infectious diseases [5]–[18]. However, as discussed above, in the context of the spread of infections (resp., misinformation), different strains of the contagion emerge as it propagates in the network due to mutations in the pathogen (resp., alteration in the information content). Therefore, it is essential to account for differing transmission risks associated with each vari-

ant/strain of the contagion. A growing body of work [4], [19]–[21] uses *multi-strain* models to analyze the spread of *mutating* contagions. Moreover, recent studies [4], [20] have highlighted that models which do not consider evolutionary adaptations can lead to incorrect predictions about the spreading dynamics of mutating contagions on contact networks.

A major challenge with existing network-based multi-strain models [4], [19], [20] is that they admit a vanishingly small clustering coefficient that tends to zero in the limit of large network size. Hence, these models cannot accurately capture some important aspects of real-world social networks, most notably the property of high clustering [16], which has a significant impact on the behavior of various spreading processes [22], [23]. To better model real-world social networks that are typically clustered, we employ a model of random networks with tuneable *clustering* and arbitrary degree distributions introduced by Miller [14] and Newman [15]. We present a summary of key references in Table I. To the best of our knowledge, our work is the first to analyze multi-strain spreading on networks with tuneable clustering. The proposed framework (Section II) enables joint evaluation of the impact of evolutionary adaptations in the contagion e.g., pathogens and misinformation, and clustering in the contact network for arbitrary degree distributions and mutation patterns.

B. Main Contributions

With the aim of understanding the mechanisms underlying epidemics caused by contagions that get altered as they propagate, we derive key epidemiological quantities for the multi-strain model [19] on random graphs with tuneable clustering [14], [15]. Our main contributions are summarized below.

i) On the theoretical side, we derive the *probability of the emergence*, i.e., the probability that a spreading process initiated by an infective seed node, selected uniformly at random, leads to an unbounded chain of infections, thus infecting a strictly positive fraction of individuals in the limit of large network size.

ii) Next, we derive the critical *epidemic threshold*, thereby defining the boundary of the region in the parameter space inside which only a finite chain of transmissions are observed with high probability and outside which epidemics occur with a positive probability.

iii) We provide extensive simulations validating our theoretical results for practical settings as well as different mutation and clustering patterns. For *doubly Poisson* contact networks, we observe that clustering increases the threshold of epidemics but reduces the probability of emergence around the phase transition point. Moreover, through an analytical case study for *one-step irreversible* mutation patterns, we observe that clustering can provide additional pathways for mutations, thereby altering the course of the epidemic.

iv) On the practical side, our results characterizing the probability of emergence and epidemic threshold pave the way for assessing the risks associated with the emergence of epidemic and information outbreaks. Our results highlight that we need to evaluate the risks of the emergence of new strains in light of policy measures that alter the structure of the contact network.

Single-strain/ Multi-strain[1]	**Clustered/ Tree-like**	**Single-layer/ Multi-layer**[2]	**Related Work**
single-strain	tree-like	single-layer	[5]–[7], [11]
single-strain	tree-like	multi-layer	[8], [12], [13]
single-strain	clustered	single-layer	[14]–[16]
single-strain	clustered	multi-layer	[17], [18]
multi-strain	*tree-like*	single-layer	[4], [19]
multi-strain	*tree-like*	multi-layer	[20]

TABLE I: *Overview of related works on network-based epidemiological models. Existing results for multi-strain spreading only focus on networks with vanishingly small clustering coefficients, while we account for multi-strain spreading in networks with tuneable clustering.*

C. Organization

We describe the clustered network and the multi-strain transmission model in Section II. In Section III, we provide our main theoretical and experimental results characterizing the emergence of epidemics. We summarize our findings and future directions in Section V. A brief outline of the proofs is provided in Section IV, with further details provided in [25].

II. Problem Setup

A. Network Model

We consider a generalization [14], [15] to the standard *configuration model* [7] that generates random graphs with arbitrary degree distribution and tuneable clustering. Note that we could quantify the level of clustering associated with a network in different ways, but here we focus on the notion of *global clustering coefficient* defined as

$$C_{\text{global}} = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triples}}$$

where a connected triple means a single vertex connected by edges to two others. The algorithm used to generate random graphs with clustering works as follows. We first specify the probability that an arbitrary node has s single-edges and is part of t triangles through the joint degree distribution $\{q_{s,t}\}_{s,t=0}^{\infty}$. Note that if a node has s single-edges and is part of t triangles, then its degree is $s + 2t$ since each triangle adds two edges connecting the node to the other end nodes of the triangle. We can view s as the number of single stubs and t as the number of corners of triangles. In order to create the network, we choose pairs of single stubs uniformly at random and join them to make a complete edge between two nodes, and we choose trios of corners of triangles at random and join them to form a triangle. The total degree distribution in the network

[1]Note that we focus on the spread of contagions where a single infectious contact can lead to infection or awareness of a piece of information, which is in contrast to models such as Linear Threshold Models [24] for complex contagions such as influence propagation.

[2]The single-layer model typically refers to a single network layer generated using the *configuration* model. In contrast, the *multi-layer* model [8], [20] is typically constructed by taking the disjoint union ($\amalg$) of network layers generated independently according to the configuration model.

is obtained through the joint distribution of single-edges and triangles $\{q_{s,t}\}_{s,t=0}^{\infty}$ as follows,

$$p_k = \sum_{s,t} q_{s,t}\delta_{k,s+2t}$$

where p_k denotes the probability that an arbitrary node is of degree k and δ_{ij} is the Kronecker delta function. In contrast to the standard configuration model, where C_{global} approaches zero in the limit of large network size, the quantity C_{global} is positive for networks generated according to the above algorithm implying the existence of a non-trivial clustering in the network. We note that the network can admit cycles comprising single-edge and triangle-edges, but they occur with a vanishingly small probability in the limit of large network size [7].

B. Transmission Model

For modeling the spread of the contagion, we adopt the *multi-strain* spreading model [19] where each *strain* or *variant* of the contagion is associated with varying risks of transmission. For $i = 1, 2, \cdots, m$, we let T_i denote the *transmissibility* of strain-i, i.e., the probability that an infectious node carrying strain-i infects its neighbor. We account for evolutionary adaptations or *mutations* in the contagion by specifying the probability μ_{ij} that strain-i mutates to strain-j within a host, where $i, j = 1, 2, \cdots, m$ and $\sum_j \mu_{ij} = 1$. The epidemiological and evolutionary processes are assumed to occur on a similar timescale, and each new infection offers an opportunity for mutation [19]. We focus on the case where $m = 2$, i.e., there are two strains propagating in the population, yet extending our theory to the general case of m strains is straightforward. The transmissibility and mutation probabilities for different strains are encoded through the transmission and mutation matrices, respectively denoted as $\boldsymbol{T}$ and $\boldsymbol{\mu}$ below.

$$\boldsymbol{T} = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{bmatrix}.$$

We consider a multi-type branching process that starts by selecting a node uniformly at random and infecting it with a particular strain, then exploring all the neighbors that are reached and infected due to this node (Figure 1). At each stage of the spreading process, a node carrying strain-i infects its neighbors independently with probability T_i. We assume that co-infection with multiple strains is not possible and subsequent to each transmission event, the contagion mutates to strain j with probability $\mu_{i,j}$, where $i, j = 1, 2, \cdots, m$ within the host. The process continues recursively until no further infections are possible. Further details regarding the multi-strain spreading setup are presented in Section III-D.

C. Metrics Studied

We characterize an outbreak as an *epidemic* if the introduction of the contagion to a host population causes an outbreak infecting a positive fraction of individuals. In contrast, we characterize outbreaks as being *self-limited* when the spreading process dies out after a finite number of transmission events. We study the following metrics [7], which are used in quantifying and assessing risks during the early stages of an outbreak.
i) The *probability of emergence* starting from strain-i is the probability that the spreading process initiated by a seed node chosen uniformly at random, carrying strain-i infects a positive fraction of the population in the limit of large network size, i.e., triggers an outbreak of size $\Omega(n)$.
ii) The *epidemic threshold* defines a boundary of the region inside which the outbreak always dies out after infecting only a finite number of individuals, while outside which an epidemic outbreak occurs with a positive probability.

III. Results and Discussion

In this section, we present our main results characterizing the probability of emergence and epidemic threshold.

A. Preliminaries

The analysis of the probability of emergence relies on recursive equations linking the number of nodes infected by the *seed* node to the number of nodes consequently infected by *later-generation* infectives. To enable such a recursive analysis, we first present preliminary facts regarding the possible configurations based on the type of strain acquired by endpoints of a triangle emanating from an infectious parent node strain-i. The situation becomes particularly challenging as compared to single-strain models [14], [15] since we need to jointly consider the status of the two nodes at the endpoints of a triangle. A graphical illustration of the resulting configurations for the case when the triangle emanates from a parent node carrying the type-1 strain (indicated in blue) is given in Figure 2. We present the corresponding probability of each configuration (p_{ij}) in Table II, with $j = 1, 2, \cdots, 6$ and where $i = 1, 2$ corresponds to the type of strain carried by the parent node. We illustrate how the probabilities are

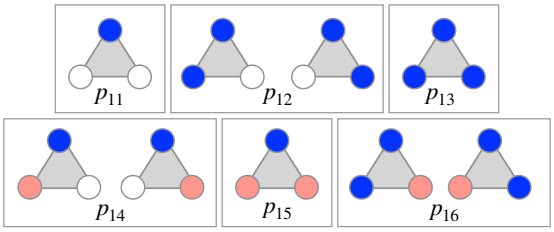

Fig. 2: *Different possible configurations for a triangle emanating from a parent node carrying strain-1. Nodes that acquire strain-1 (resp., strain-2) after mutation are indicated in blue (resp., red). The configurations are based on whether the node at either endpoint of the triangle gets infected and the resulting strain it acquires after mutation.*

derived for a sample configuration p_{13} in Figure 2 and direct the reader to [25] for derivations of the remaining scenarios in Table II. In Figure 2, the configuration corresponding to p_{13} occurs when i) the parent node infects both endpoints of the triangle, and they acquire strain-1, *or* ii) the parent node infects one of the two endpoints (say the left node) but fails to infect the other endpoint (say the right node) which later gets infected (due to the left node). Hence, the probability for this

p_{i1}	$(1-T_i)^2$
p_{i2}	$2T_i\mu_{i1}(1-T_i)(1-T_1)$
p_{i3}	$(T_i\mu_{i1})^2 + 2T_i\mu_{i1}(1-T_i)T_1\mu_{11}$
p_{i4}	$2T_i\mu_{i2}(1-T_i)(1-T_2)$
p_{i5}	$(T_i\mu_{i2})^2 + 2T_i\mu_{i2}(1-T_i)T_2\mu_{22}$
p_{i6}	$2\left(T_i^2\mu_{i1}\mu_{i2} + T_i\mu_{i1}(1-T_i)T_1\mu_{12} + T_i\mu_{i2}(1-T_i)T_2\mu_{21}\right)$

TABLE II: *The probability p_{ij} of occurrence for each of the scenarios in Figure 2, where $i = 1, 2$ corresponds to the strain carried by the parent and $j = 1, \ldots, 6$ corresponds to the configuration at the endpoints of the triangle emanating at the parent node.*

configuration is $(T_1\mu_{11})^2 + 2T_1\mu_{11}(1-T_1)T_1\mu_{11}$, where the factor 2 is due to symmetry. Now that we have established the framework for presenting our main results, we present our first analytical result characterizing the probability of emergence.

B. Probability of Emergence

Theorem 3.1 (Probability of Emergence): For multi-strain spreading with parameters $(\boldsymbol{T}, \boldsymbol{\mu})$, initiated by a randomly selected seed node carrying strain-i, on a clustered network with a given joint degree distribution of single-edges and triangles $(q_{s,t})$, for $i = 1, 2$, we have

$$\mathbb{P}[\text{Emergence}] = 1 - \sum_{s,t} q_{s,t}(h_i(1))^s(g_i(1))^t, \tag{1}$$

where $h_i(1), g_i(1)$ are the smallest non-negative roots of the fixed point equations:

$$\begin{aligned} h_i(1) &= 1 - T_i + T_i\Big(\mu_{i1}\sum_{s,t}\frac{sq_{s,t}}{\langle s\rangle}h_1(1)^{s-1}g_1(1)^t \\ &\quad + \mu_{i2}\sum_{s,t}\frac{sq_{s,t}}{\langle s\rangle}h_2(1)^{s-1}g_2(1)^t\Big), \end{aligned} \tag{2}$$

$$\begin{aligned} g_i(1) &= p_{i1} + p_{i2}\left(\sum_{s,t}\frac{tq_{s,t}}{\langle t\rangle}h_1(1)^s g_1(1)^{t-1}\right) \\ &\quad + p_{i3}\left(\sum_{s,t}\frac{tq_{s,t}}{\langle t\rangle}h_1(1)^s g_1(1)^{t-1}\right)^2 \\ &\quad + p_{i4}\left(\sum_{s,t}\frac{tq_{s,t}}{\langle t\rangle}h_2(1)^s g_2(1)^{t-1}\right) \\ &\quad + p_{i5}\left(\sum_{s,t}\frac{tq_{s,t}}{\langle t\rangle}h_2(1)^s g_2(1)^{t-1}\right)^2 \\ &\quad + p_{i6}\left(\sum_{s,t}\frac{tq_{s,t}}{\langle t\rangle}h_1(1)^s g_1(1)^{t-1}\right) \\ &\quad \cdot\left(\sum_{s,t}\frac{tq_{s,t}}{\langle t\rangle}h_2(1)^s g_2(1)^{t-1}\right), \ i = 1, 2. \end{aligned} \tag{3}$$

We provide an outline for the proof of Theorem 3.1 in Section IV-A and direct the reader to [25] for more details. We present numerical simulations in Section III-D.

C. Epidemic Threshold

Our following result characterizes the epidemic threshold.

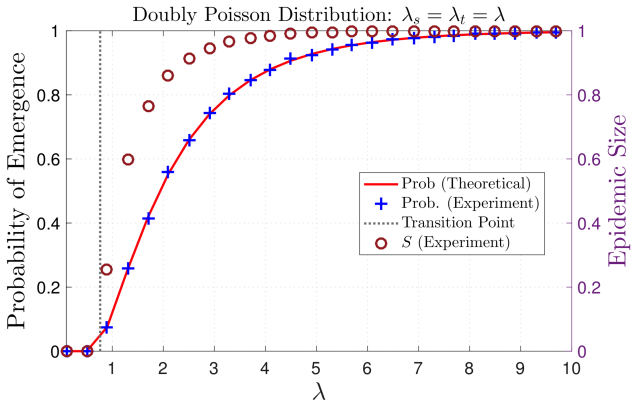

Fig. 3: *The probability of emergence on contact networks with doubly Poisson distribution* (5)*, with the distribution for single-edges and triangles, respectively parameterized by λ_s and λ_t. The theoretical probability and transition points are derived from Theorems 3.1 and 3.2, respectively. The experimental probability of emergence is obtained by averaging over 1.5×10^4 experiments. We also plot the conditional mean for the outbreak size (S), given that it occurs, as observed in the experiments. The network size n is 2×10^5 and the number of independent experiments for data point is 1.5×10^4 for $\lambda_s = \lambda_t = \lambda$ and we vary λ in the interval $(0, 10)$.*

Theorem 3.2 (Epidemic Threshold): For multi-strain spreading with parameters $(\boldsymbol{T}, \boldsymbol{\mu})$ on a clustered network, we define

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{\Pi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta} \end{bmatrix} \begin{bmatrix} \frac{\langle s^2\rangle - \langle s\rangle}{\langle s\rangle}\mathbf{I} & \frac{\langle st\rangle}{\langle s\rangle}\mathbf{I} \\ \frac{\langle st\rangle}{\langle t\rangle}\mathbf{I} & \frac{\langle t^2\rangle - \langle t\rangle}{\langle t\rangle}\mathbf{I} \end{bmatrix}. \tag{4}$$

Let $\sigma(\boldsymbol{J})$ denote the spectral radius of $\boldsymbol{J}$. The epidemic threshold is given by $\sigma(\boldsymbol{J}) = 1$, where

$$\boldsymbol{\Pi} = \begin{bmatrix} T_1\mu_{11} & T_1\mu_{12} \\ T_2\mu_{21} & T_2\mu_{22} \end{bmatrix},$$

$$\boldsymbol{\Delta} = \begin{bmatrix} p_{12} + 2p_{13} + p_{16} & p_{14} + 2p_{15} + p_{16} \\ p_{22} + 2p_{23} + p_{26} & p_{24} + 2p_{25} + p_{26} \end{bmatrix}.$$

Observe that for $i, j = 1, 2$, the matrix entry $\boldsymbol{\Pi}_{ij}$ corresponds to the probability that a parent node carrying strain-i infects its neighbor via a single-edge with strain-j after mutation. Similarly, $\boldsymbol{\Delta}_{ij}$ corresponds to the mean number of nodes that acquire strain-j at the endpoints of a triangle emanating from a parent node carrying strain-i. We note that the epidemic threshold as presented in Theorem 3.2 is a strict generalization of the multi-strain model without clustering, the threshold for which can be inferred by substituting $\boldsymbol{\Delta} = 0$ in (4). In Section III-E we further decompose the epidemic threshold to delineate the impact of clustering through a case-study with irreversible mutations. A brief outline for proving Theorem 3.2 is provided in Section IV-B; the full derivation is available in [25].

D. Simulations

Next, we describe the simulation setup and results. Unless stated otherwise, the spreading process is initiated by selecting a *seed* node uniformly at random and infecting it with strain-1. The seed node infects each neighbor independently with probability T_1, following which, each newly infected neighbor

mutates independently to strain-2 with probability μ_{12}. In the k^{th} round, a node that was infected by a node carrying strain-i first undergoes mutation with probability given through the mutation matrix before attempting to infect its neighbors during the $k{+}1^{\text{th}}$ round. Each node is assumed to be infectious only for one round. As the infections continue to grow, both strains might co-exist in the population. Moreover, the presence of cycles in the contact network can simultaneously expose a susceptible node to multiple infections. We assume that co-infection is not possible, and resolve the exposure to multiple infections as follows. If a node is exposed to x infections of strain-1 and y infections of strain-2 simultaneously, the node becomes infected with strain-1 (respectively, strain-2) with probability $x/(x+y)$ (respectively, $y/(x+y)$) for any non-negative constants x and y. The process terminates when no further infections are possible.

While our results hold for *arbitrary* distributions for the triangles and single-edge, as a concrete illustration we consider the setting where the joint degree sequence $q_{s,t}$ is given by the *doubly Poisson distribution*, i.e., the number of single-edges and triangles are independent, and they follow a Poisson distribution. Namely, we set

$$q_{s,t} = e^{-\lambda_s}\frac{(\lambda_s)^s}{s!}\cdot e^{-\lambda_t}\frac{(\lambda_t)^t}{t!}, \quad s,t = 1,\dots \tag{5}$$

with λ_s and λ_t denoting the mean number of single-edges and triangles, respectively. For the experiments in Figures 3 and 4, we set $T_1 = 0.2, T_2 = 0.5, \mu_{11} = \mu_{22} = 0.75$.

In Figure 3, we consider the cases when $\lambda_s = \lambda_t = \lambda$ with λ varying from 1 to 10. For each value of λ, we obtain the empirical probability of emergence. In particular, we set the network size n to 2×10^5 and perform 1.5×10^4 independent experiments for each data point. The empirical probability of emergence is given by the fraction of experiments for which an outbreak emerges. In addition, we compute the critical value of λ for which (4) has a spectral radius of one, i.e., $\sigma(\boldsymbol{J}) = 1$, to mark the phase transition point. We also plot the expected epidemic size S obtained by the simulations. We observe that both the probability of emergence and the epidemic size transition from zero to a positive value around the epidemic threshold. Our theoretical results on the probability of emergence and phase transition point are in agreement with simulation results.

Next, to evaluate the impact of clustering, we consider a joint degree distribution that allows us to control the level of clustering, while keeping the mean total degree fixed. In particular, for each node, we set the number of incident single-edges as $2\times\text{Poisson}\left(\frac{4-c}{2}\lambda\right)$ and the number of triangles to $\text{Poisson}\left(\frac{c}{2}\lambda\right)$ where $c\in[0,4]$. This ensures that as c varies, both the mean and the variance of the degree distribution remains constant, allowing us to focus only on the effect of clustering. The parameter c controls the level of clustering in the contact network. As c increases, the clustering coefficient of the network also increases. Observe that when $c = 0$, there will be no triangles in the network, and the clustering coefficient will be close to zero. In contrast, when $c = 4$, there will be no single-edges in the network, and the clustering coefficient will be close to one. As c increases, the clustering coefficient of the network also increases.

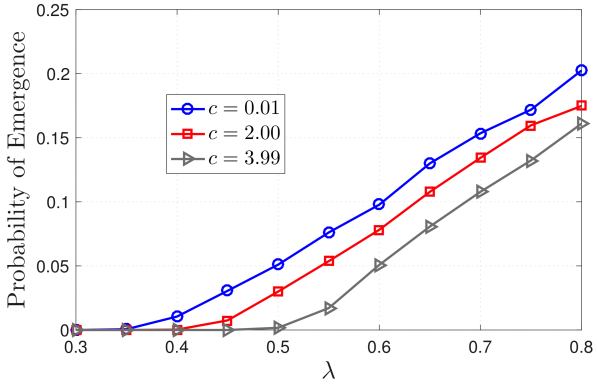

Fig. 4: *The impact of clustering: For each node, the number of incident single-edges is* $2\times\text{Poi}\left(\frac{4-c}{2}\lambda\right)$ *and the number of triangles is* $\text{Poi}\left(\frac{c}{2}\lambda\right)$. *This ensures varying* c *keeps the mean degree and excess degree distribution fixed while changing the clustering coefficient. The network size* n *is* 2×10^5, *and the number of independent experiments for each data point is* 10^4. *Our experimental results show that high clustering increases the threshold of epidemics and reduces the probability of emergence around the transition point.*

In Figure 4, we consider three different values for the parameter c, namely, $c = 0.01$, $c = 2.00$, and $c = 3.99$, respectively to illustrate the impact of the clustering coefficient on the probability of emergence and the epidemic threshold. Our results reveal that high clustering increases the threshold of epidemics and reduces the probability of emergence around the transition point. These observations are consistent with the single-strain spreading [17] on clustered networks.

E. Joint Impact of Clustering and Evolution

Next, we discuss the interplay of clustering and evolution on the probability of emergence of epidemic outbreaks. We consider the case where the fitness landscape consists of two strains. The process starts when the population is introduced to the first strain (strain-1) which is moderately transmissible and initially dominant in the population. In contrast, the other strain (strain-2) is highly transmissible and initially absent in the population but has the risk of emerging through mutations in strain-1. For $\mu_{22} = 1$ and $\mu_{12}\in[0,1]$, we have:

$$\boldsymbol{\mu} = \begin{bmatrix} 1-\mu_{12} & \mu_{12} \\ 0 & 1 \end{bmatrix}; \quad \boldsymbol{T} = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix}, \quad T_1 < T_2. \tag{6}$$

The above mutation and transmission parameters (6) correspond to the *one-step irreversible* mutation scheme, which is used widely [19], [21] to model scenarios where a simple change is required for the contagion to evolve to a highly transmissible variant. We first isolate the impact of clustering in altering the epidemic threshold in the following Lemma.

Lemma 3.3: For multi-strain spreading with one-step irreversible mutations (6) *on clustered networks with doubly Poisson distribution* (5), *the epidemic threshold is given as:*

$$\sigma(\boldsymbol{J}) = \lambda_s T_2 \times \left(1 + \left(\frac{2\lambda_t}{\lambda_s}(1 - T_2^2 + T_2)\right)\right). \tag{7}$$

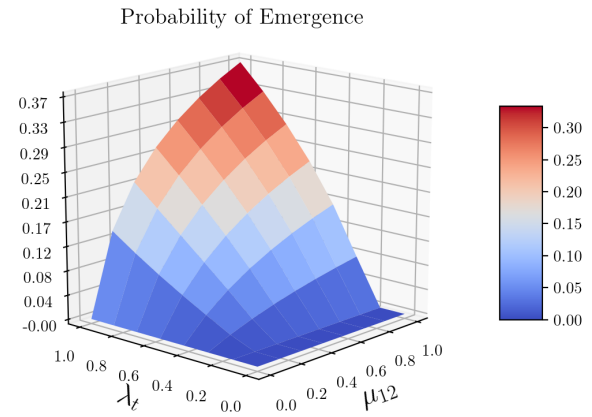

Fig. 5: *We consider a contact network following a doubly Poisson distribution* (5) *and with one-step irreversible mutations* (6)*. we plot the probability of emergence given by Theorem 3.1 as a function of* (λ_t, μ_{12})*. Here,* $\lambda_t = 0$ *and* $\mu_{12} = 0$ *respectively correspond to the absence of clustering and mutations. We set* $T_1 = 0.2, T_2 = 0.7, \mu_{22} = 1, \lambda_s = 1$*. We observe that the joint impact of clustering and evolution* $(\lambda_t > 0, \mu_{12} > 0)$ *can lead to a significant increase in the probability of emergence of an epidemic as compared to the case when either of these phenomena act in isolation* ($\lambda_t = 0$ *or* $\mu_{12} = 0$)*.*

Through (7), we can see that compared to a network with only single-edges (distributed as $\text{Poisson}(\lambda_s)$), the epidemic threshold increases as a multiplicative factor of $\left(1 + \left(\frac{2\lambda_t}{\lambda_s}(1 - T_2^2 + T_2)\right)\right)$. Moreover, when $\lambda_t = 0$, we can recover the epidemic threshold of $\sigma(\boldsymbol{J}) = \lambda_s T_2$ corresponding to one-step irreversible mutations on a network with a vanishingly small clustering coefficient [4], [19]. A derivation of Lemma 3.3 is presented in [25].

In what follows, we note that the addition of clustering in the network structure can offer additional pathways for mutations which can, in turn, alter the course of the pandemic. We demonstrate this phenomenon for the case when the distribution of the single-edges and triangles is doubly Poisson with parameters λ_s and λ_t respectively, as in (5). We vary λ_t and μ_{12} in the interval $[0, 1]$ to parameterize the transition from a single-strain to a multi-strain spreading and from an unclustered to a clustered network. Note that when $\lambda_t = 0$, the network only comprises of single-edges and has a vanishingly small clustering coefficient [7]. On the other hand, since the spreading process is initiated with strain-1, setting $\mu_{12} = 0$ corresponds to a single-strain setting.

In Figure 5, we invoke Theorem 3.1 and plot the probability of emergence as a function of (λ_t, μ_{12}), while setting $T_1 = 0.2, T_2 = 0.7, \mu_{22} = 1$, and $\lambda_s = 1$. We observe that the probability of emergence remains low when either $\mu_{12} = 0$ *or* $\lambda_t = 0$, i.e., the likelihood of seeing an epidemic remains small when clustering or mutations act in isolation. In other words, even when mutations occur with a positive probability, the absence of triangles $(\lambda_t = 0)$ renders a negligible risk of an epidemic outbreak. Similarly, in the absence of mutations $(\mu_{12} = 0)$, clustering alone does not lead to an increased risk of an epidemic. However, for $\mu_{12} = \lambda_t = 1$, i.e., when mutations occur with a high probability *and* clustering is significant, we observe the probability of emergence rises sharply (Figure 5). This observation further highlights the need for evaluating risks of emergence of highly contagious mutations in the light of the structure of the contact network.

IV. Proof Sketch

A. Outline for Proof of Theorem 3.1

To see intuitively why Theorem 3.1 holds, we note that $h_i(x)$ (respectively, $g_i(x)$) corresponds to the probability generating function (PGF) of the number of *finite* nodes reached and infected by following a randomly selected single-edge (respectively, triangle) emanating from a node carrying strain-i. This gives a way to define the PGF of the number of finite nodes reached and infected by selecting a node uniformly at random and making it type-i, denoted by $Q_i(x)$. Observe that

$$Q_i(x) = x \sum_{s,t} q_{s,t} h_i(x)^s g_i(t)^t. \tag{8}$$

The validity of (8) can be seen as follows– the factor x accounts for the node which is selected randomly and given the infection as the seed of the process. Note that this node has a joint degree (s, t) with probability $q_{s,t}$. Since this node carries strain-i, the number of nodes reached and infected by each of its s single-edges (respectively, each of the t triangles) has a generating function $h_i(x)$ (respectively, $g_i(x)$). From the *powers* property of PGFs [26], the total number of nodes reached and infected in this process when the initial node carries strain-i and has joint degree (s, t) has a generating function $h_i(x)^s g_i(x)^t$. As we average over all possible joint degrees (s, t), we obtain (8). Note that when $Q(i) = 1$, the number of infected nodes is *finite* from the conservation of probability. Whereas when $Q(i) < 1$, the corresponding probability $1 - Q(i)$ gives the probability of infecting an *infinite* number of nodes leading to an epidemic outbreak.

Next, we observe that for a node reached by following a single-edge (resp., triangle) selected uniformly at random, the joint degree distribution is proportional to the number of single-edges (resp., triangle) assigned to that node and given by $sq_{s,t}/\langle s \rangle$ (resp., $tq_{s,t}/\langle t \rangle$) where $\langle s \rangle = \sum_{s,t} sq_{s,t}$ (resp., $\langle t \rangle = \sum_{s,t} tq_{s,t}$) ensures normalization. We first state and explain how to derive $h_1(x)$, the PGF of the number of finite nodes reached and infected by following a randomly selected single-edge emanating from a node carrying strain-1;

$$h_1(x) = 1 - T_1 + T_1 x \Big(\mu_{11} \sum_{s,t} \frac{sq_{s,t}}{\langle s \rangle} h_1(x)^{s-1} g_1(x)^t + \mu_{12} \sum_{s,t} \frac{sq_{s,t}}{\langle s \rangle} h_2(x)^{s-1} g_2(x)^t \Big). \tag{9}$$

We note that if no transmission occurs along the randomly selected single-edge emanating from a node carrying strain-1, we get the factor of $(1 - T_1)x^0$ in (9). Whereas, if transmission occurs along the selected single-edge, the number of nodes eventually infected will be one *plus* all the nodes reached and infected due to node at the endpoint of the selected edge, which contributes a factor $T_1 x \Big(\mu_{11} \sum_{s,t} \frac{sq_{s,t}}{\langle s \rangle} h_1(x)^{s-1} g_1(x)^t + \mu_{12} \sum_{s,t} \frac{sq_{s,t}}{\langle s \rangle} h_2(x)^{s-1} g_2(x)^t \Big)$. This follows from noting that

the node reached by following the randomly selected single-edge has already utilized one of its single-edges to connect to its parent, and it has $s-1$ remaining single-edges and t triangles. Moreover, this node acquires strain-1 (resp., strain-2) with probability μ_{11} (resp., μ_{12}). When this node carries strain-1 (resp., strain-2), the powers property of PGFs readily implies that the number of nodes infected due to this node has a generating function $h_1(x)^{s-1}g_1(x)^t$ (resp., $h_2(x)^{s-1}g_2(x)^t$). Averaging over all possible joint degrees and node types yields (9). Due to space constraints, we direct the reader to [25] for the derivation for the PGFs $h_1(x)$, $h_2(x)$, $g_1(x)$, and $g_2(x)$.

B. Outline for Proof of Theorem 3.2

The proof of Theorem 3.2 is linked to the stability of the fixed point solutions of (2, 3). We note that for $i = 1, 2$, the set of equations (2, 3) admits a trivial fixed point $h_1(1) = h_2(1) = g_1(1) = g_2(1) = 1$. Substituting back into (8) gives $1 - Q_i(1) = 0$, i.e., all infected components are of finite size, and no outbreak emerges. To check the stability of this trivial solution, we linearize the set of equations (2, 3) around $h_1(1) = h_2(1) = g_1(1) = g_2(1) = 1$ and compute the corresponding Jacobian matrix $\boldsymbol{J} = [J_{ij}]$. If $\sigma(\boldsymbol{J}) \leq 1$, then the trivial solution is stable, leading to a zero probability of emergence. However, if $\sigma(\boldsymbol{J}) > 1$, then there exists another stable solution with $h_1(1), h_2(1), g_1(1), g_2(1) < 1$, leading to a positive probability of emergence, i.e., $1 - Q_i(1) > 0$. Therefore, $\sigma(\boldsymbol{J}) = 1$ emerges as the epidemic threshold.

V. Conclusions

We analyzed multi-strain spreading on networks with tuneable clustering and arbitrary degree distributions. We derived the probability of emergence of an epidemic outbreak and the critical epidemic threshold beyond which epidemics occur with a positive probability. Our framework allows for evaluating the emergence of highly transmissible variants in relation to the extent of clustering in the network. Future directions include leveraging the multi-strain model on social network data to gain insights into the efficacy of countermeasures to combat the unwarranted spread of misinformation.

Acknowledgements

This research was supported in part by the National Science Foundation through grants 2225513, 2026985 and 1813637 and by the Army Research Office through grant #W911NF-22-1-0181. O.Y. also acknowledges the IBM academic award. M.S. acknowledges the Dowd Fellowship, Knight Fellowship, Lee-Stanziale Ohana Endowed Fellowship, and Cylab Presidential Fellowship from Carnegie Mellon University.

References

[1] W. T. Harvey, A. M. Carabelli, B. Jackson, R. K. Gupta, E. C. Thomson, E. M. Harrison, C. Ludden, R. Reeve, A. Rambaut, S. J. Peacock *et al.*, "Sars-cov-2 variants, spike mutations and immune escape," *Nature Reviews Microbiology*, vol. 19, no. 7, pp. 409–424, 2021.

[2] O. Papakyriakopoulos, J. C. M. Serrano, and S. Hegelich, "The spread of covid-19 conspiracy theories on social media and the effect of content moderation," *The Harvard Kennedy School (HKS) Misinformation Review*, vol. 18, 2020.

[3] L. A. Adamic, T. M. Lento, E. Adar, and P. C. Ng, "Information evolution in social networks," in *ACM WSDM*, 2016, pp. 473–482. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835827

[4] R. Eletreby, Y. Zhuang, K. M. Carley, O. Yağan, and H. V. Poor, "The effects of evolutionary adaptations on spreading processes in complex networks," *Proceedings of the National Academy of Sciences*, vol. 117, no. 11, pp. 5664–5670, 2020. [Online]. Available: https://www.pnas.org/content/117/11/5664

[5] M. E. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E*, vol. 66, p. 016128, Jul 2002. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.66.016128

[6] E. Kenah and J. M. Robins, "Second look at the spread of epidemics on networks," *Phys. Rev. E*, vol. 76, no. 3, p. 036113, 2007.

[7] M. E. Newman, *Networks*. Oxford university press, 2018.

[8] O. Yağan, D. Qian, J. Zhang, and D. Cochran, "Conjoining speeds up information diffusion in overlaying social-physical networks," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 6, pp. 1038–1048, 2013.

[9] O. Yağan, D. Qian, J. Zhang, and D. Cochran, "Information diffusion in overlaying social-physical networks," in *2012 46th Annual Conference on Information Sciences and Systems (CISS)*. IEEE, 2012, pp. 1–6.

[10] Y. Tian, A. Sridhar, O. Yağan, and H. V. Poor, "Analysis of the impact of mask-wearing in viral spread: Implications for covid-19," in *2021 American Control Conference (ACC)*. IEEE, 2021, pp. 3132–3137.

[11] Y. Tian, A. Sridhar, H. V. Poor, and O. Yagan, "The role of masks in mitigating viral spread on networks," *arXiv preprint arXiv:2110.04398*, 2021.

[12] A. Hackett, D. Cellai, S. Gómez, A. Arenas, and J. P. Gleeson, "Bond percolation on multiplex networks," *Phys. Rev. X*, vol. 6, p. 021002, Apr 2016. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.6.021002

[13] G. Bianconi, "Epidemic spreading and bond percolation on multilayer networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2017, no. 3, p. 034001, mar 2017. [Online]. Available: https://dx.doi.org/10.1088/1742-5468/aa5fd8

[14] J. C. Miller, "Percolation and epidemics in random clustered networks," *Physical Review E*, vol. 80, no. 2, p. 020901, 2009.

[15] M. E. Newman, "Random graphs with clustering," *Physical review letters*, vol. 103, no. 5, p. 058701, 2009.

[16] M. Á. Serrano and M. Boguna, "Clustering in complex networks. i. general formalism," *Physical Review E*, vol. 74, no. 5, p. 056114, 2006.

[17] Y. Zhuang and O. Yağan, "Information propagation in clustered multi-layer networks," *IEEE Transactions on Network Science and Engineering*, vol. 3, no. 4, pp. 211–224, 2016.

[18] P. Mann, V. A. Smith, J. B. O. Mitchell, and S. Dobson, "Random graphs with arbitrary clustering and their applications," *Phys. Rev. E*, vol. 103, p. 012309, Jan 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.103.012309

[19] H. Alexander and T. Day, "Risk factors for the evolutionary emergence of pathogens," *Journal of The Royal Society Interface*, vol. 7, no. 51, pp. 1455–1474, 2010.

[20] M. Sood, A. Sridhar, R. Eletreby, C. W. Wu, S. A. Levin, H. V. Poor, and O. Yağan, "Spreading processes with mutations over multi-layer networks," 2022. [Online]. Available: https://arxiv.org/abs/2210.05051

[21] R. Antia, R. R. Regoes, J. C. Koella, and C. T. Bergstrom, "The role of evolution in the emergence of infectious diseases," *Nature*, vol. 426, no. 6967, p. 658, 2003.

[22] A. Hackett, S. Melnik, and J. P. Gleeson, "Cascades on a class of clustered random networks," *Physical Review E*, vol. 83, no. 5, p. 056107, 2011.

[23] X. Huang, S. Shao, H. Wang, S. V. Buldyrev, H. E. Stanley, and S. Havlin, "The robustness of interdependent clustered networks," *EPL (Europhysics Letters)*, vol. 101, no. 1, p. 18002, 2013.

[24] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *2010 IEEE International Conference on Data Mining*, 2010, pp. 88–97.

[25] M. Sood, R. Eletreby, S. Kumar, and O. Yağan, "The interplay of clustering and evolution in the emergence of epidemics on networks." [Online]. Available: http://users.ece.cmu.edu/~msood/preprints/icc22.pdf

[26] M. E. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distributions and their applications," *Physical review E*, vol. 64, no. 2, p. 026118, 2001.