Measuring event concentration in empirical networks with different types of degree distributions

Juan Campos, Jorge Finke

1 Sports Models Division, Genius Sports, Medellin, Colombia, 2 Department of Electrical Engineering and Computer Science, Pontificia Universidad Javeriana, Cali, Colombia

Abstract

Measuring event concentration often involves identifying clusters of events at various scales of resolution and across different regions. In the context of a city, for example, clusters may be characterized by the proximity of events in the metric space. However, events may also occur over urban structures such as public transportation and infrastructure systems, which are naturally represented as networks. Our work provides a theoretical framework to determine whether events distributed over a set of interconnected nodes are concentrated on a particular subset. Our main analysis shows how the proposed or any other measure of event concentration on a network must explicitly take into account its degree distribution.

We apply the framework to measure event concentration (i) on a street network (i.e., approximated as a regular network where events represent criminal activities); and (ii) on a social network (i.e., a power law network where events represent users who are dissatisfied after purchasing the same product).

Introduction

Consider a non-uniformly distribution of events over different regions. Past efforts to explain the mechanisms through which some regions reveal a high concentration of events (i.e., form hotspots) range from agent-based [1], game theoretic [2, 3], reaction-diffusion [4], and predator-prey [5] modeling. Generally, these approaches account for the relative location of events on the metric space, and use kernel density techniques to identify and recreate hotspots [6, 7]. For a number of scenarios, however, the concentration of events can be better captured by their distribution over a set of interconnected nodes [8, 9].

The work in [8] adapts the main idea behind kernel density techniques [6, 7] to identify hotspots on networks. In particular, a hotspot indicates a subnetwork that contains the maximum number of events on the smallest total path length. The authors consider two types of networks, namely, binary trees and regular networks. To compute the optimal subnetwork for binary trees, they introduce an algorithm that identifies hotspots based on dynamic programming. For regular networks, identifying hotspots requires that all possible subnetworks be
Measuring event concentration in empirical networks

Consider an undirected network \( G = (V, E) \), where \( V = \{v_1, \ldots, v_n\} \) denotes the set of nodes and \( E \subseteq V \times V \) the set of edges. The geodesic distance between nodes \( v_i \) and \( v_j \) is denoted by \( \rho(v_i, v_j) \). We borrow the definition of a Voronoi diagram of a network from [11].

**Definition 1.** Suppose that there exists a subset of nodes marked as generator nodes. This subset is denoted by \( U = \{u_1, \ldots, u_m\} \subseteq V \). The Voronoi diagram of \( G = (V, E) \) associated to the nodes belonging to \( U \) is a partition \( \{V(u_1), \ldots, V(u_m)\} \) of \( V \), such that:

- If \( v_i \in V(u_s) \), then \( \rho(v_i, u_s) \leq \rho(v_i, u_{s'}) \) for all \( s' \in \{1, \ldots, m\} \).
- If \( \rho(v_i, u_s) = \rho(v_i, u_{s'}) \), then node \( v_i \) is equally likely to be assigned to \( V(u_s) \) or \( V(u_{s'}) \).

Let \( u_i \) represent a generator node which is associated to the occurrence of at least a certain number of events \( e_i \geq 0 \). A generator node may capture, for example, an intersection on a representation of a street network where \( e \) or more criminal activities occur. Regular nodes, on the other hand, represent intersections at which less than \( e \) events occurs. Such nodes belong to the set \( U^c = V - U \). Note that the generator node associated to cell \( V(u_i) \) is denoted by \( u_i \) and a cell refers to an element of the Voronoi partition. Note also that any cell \( V(u_i) \) contains one generator node. Finally, note that if the network \( G \) is a connected network, then any regular node \( v_i \in V \) belongs to some cell.

evaluated, which becomes computationally costly for networks of large size. A second shortcoming of the work in [8] is that the approach does not extend to networks with more realistic topologies. In most cases, empirical networks exhibit degree distributions under which the nodes connect to different numbers of neighbors (ranging orders of magnitude in degree values, e.g., for networks with power law degree distributions).

The work in [9] introduces an alternative approach, which evaluates the concentration of events on networks with different types of degree distributions (namely, networks with with regular, Poisson, and power law degree distributions) based on Voronoi diagrams [10]. Nodes that are associated with the occurrence of a certain number of events are marked as *generator nodes*. Voronoi cells are then defined according the geodesic distances from generator nodes to all other nodes of the network. The measure of concentration of events that the authors propose builds on a key property of Voronoi diagrams: groups of small, adjacent cells (created by generator nodes) correspond to subnetworks with a high event concentration.

Simulation results in [9] illustrate that evaluating event concentration in networks depends on their degree distribution. This paper extends the work in [9] in two ways. First, we provide the mathematical foundation to characterize the theoretical distribution of the sizes of the Voronoi cells when events are located uniformly at random over a network with an arbitrary distribution. Second, we use the resulting distribution and apply the criterion in [9] to measure event concentrations in empirical networks. In particular, we consider events that represent (i) criminal activities on a street network (approximated as a regular network), and (ii) dissatisfied users on a social network (a power law network). Our results illustrate the importance of understanding the relationship between the degree distribution and the dispersion of events on a network in order to identify and recreate the formation of hotspots.

The remaining sections are organized as follows. Section 2 introduces some notation and preliminaries. Section 3 presents the proposed framework for measuring event concentration, and introduces the criterion for detecting hotspots based on a summary statistic derived from the proposed framework. Section 4 applies the criterion to the two empirical networks and compares the outcome to that of detecting hotspots by the proximity of events in the metric space. Finally, Section 5 draws some conclusions and future research directions.

**Preliminaries**

Consider an undirected network \( G = (V, E) \), where \( V = \{v_1, \ldots, v_n\} \) denotes the set of nodes and \( E \subseteq V \times V \) the set of edges. The geodesic distance between nodes \( v_i \) and \( v_j \) is denoted by \( \rho(v_i, v_j) \). We borrow the definition of a Voronoi diagram of a network from [11].

**Definition 1.** Suppose that there exists a subset of nodes marked as generator nodes. This subset is denoted by \( U = \{u_1, \ldots, u_m\} \subseteq V \). The Voronoi diagram of \( G = (V, E) \) associated to the nodes belonging to \( U \) is a partition \( \{V(u_1), \ldots, V(u_m)\} \) of \( V \), such that:

- If \( v_i \in V(u_s) \), then \( \rho(v_i, u_s) \leq \rho(v_i, u_{s'}) \) for all \( s' \in \{1, \ldots, m\} \).
- If \( \rho(v_i, u_s) = \rho(v_i, u_{s'}) \), then node \( v_i \) is equally likely to be assigned to \( V(u_s) \) or \( V(u_{s'}) \).

Let \( u_i \) represent a generator node which is associated to the occurrence of at least a certain number of events \( e_i \geq 0 \). A generator node may capture, for example, an intersection on a representation of a street network where \( e \) or more criminal activities occur. Regular nodes, on the other hand, represent intersections at which less than \( e \) events occurs. Such nodes belong to the set \( U^c = V - U \). Note that the generator node associated to cell \( V(u_i) \) is denoted by \( u_i \) and a cell refers to an element of the Voronoi partition. Note also that any cell \( V(u_i) \) contains one generator node. Finally, note that if the network \( G \) is a connected network, then any regular node \( v_i \in V \) belongs to some cell.
Based on Definition 1, \( n_s = |V(u_s)| \geq 1 \) denotes the size of cell \( V(u_s) \). The distribution of \( n_s \) for all \( u_s \in U \) for \( G \), that is, the distribution of the sizes of all cells determines whether events on \( G = (V, E) \) are uniformly distributed. Deviations from a uniform distribution evidence a concentration of events that results from a non-uniform allocation.

Fig 1 shows the Voronoi cells in a scenario with two particular cases. Note that in Fig 1(a) the non-uniform allocation of the generator nodes yields relative small geodesic distances between them. In comparison to Fig 1(b), where generator nodes (events) are uniformly distributed, most cells of the Voronoi diagram in Fig 1(a) contain a small number of regular nodes. Fig 2 depicts the probability mass function (pmf) of the sizes of the cells for uniform and non-uniform event allocations.

Next, consider a randomly selected node \( v \), with degree \( d_v = d \). Let \( N_v^\delta = \{ v' \in V : \rho(v, v') = \delta \} \) represent the neighborhood of nodes located at a distance \( \delta \) from node \( v \). Furthermore, let \( D \) denote a random variable that represents the degree of a randomly selected node. And finally, let \( D_v^\delta \) denote a random variable that represents the degree of a randomly selected node in \( N_v^\delta \).

To derive the pmf of the sizes of the Voronoi cells resulting from a uniform distribution of events, consider the following assumptions. Suppose that the degree distribution (i.e., the pmf of \( D \)) and the conditional degree distribution (i.e., the pmf of \( D_v^\delta \)) are known [12]. Furthermore, suppose that the pmf of \( D_v^\delta \) for \( \delta \geq 2 \) can be approximated as

\[
P[D_v^\delta = d_v^\delta] \approx \frac{nP[D = d_v^\delta] d_v^\delta}{2|E|} = \frac{P[D = d_v] d_v}{k}
\]

where \( k \) is the average degree of all nodes of the network.

Under the above assumptions, we are now able to introduce a framework that defines the pmf of the sizes of Voronoi cells when events are distributed uniformly at random. The notation for the framework is summarized in Table 1.
Framework and hotspot detection criterion

Let $D_g$ denote a random variable that represents the degree of a randomly selected generator node. Furthermore, the proportion of nodes that are generator nodes is denoted by $p = m/n$, $0 < p < 1$. Consider a randomly selected generator node, denoted node $u$. For convenience, we will write $N_u$ to denote $N_u^0$.

**Assumption 1.** Suppose that

1. The degree distribution of the generator nodes resembles the degree distribution of all nodes of $G$.
2. If $v \in V(u), v \neq u$, then $v \in N_1 \cup N_2$.
3. The local clustering coefficient of node $u$ is negligible (less than 0.1).
4. Nodes in $N_2$ have a single neighbor in $N_1$.

Fig 3 illustrates conditions 2-4 of Assumption 1. These conditions are satisfied for Fig 3(a) but not for Fig 3(b).

The degree of generator node $u$ is denoted by $d_u = d$. Note that the random variable $D_g^0$ represents the degree of a randomly selected node in $N_g$. Let $X_d$ denote a random variable that

| Notation | Meaning |
|----------|---------|
| $X$ | Size of the cell of a randomly selected generator node |
| $X_d$ | Size of the cell of a randomly selected generator node $u$ with $d_u = d$ |
| $X_1^d$ | Number of nodes in $V(u) \cap N_u^1$, where $d_u = d$ |
| $X_2^d$ | Number of nodes in $V(u) \cap N_u^2$, $d_u = d$, that are themselves neighbors of node $v_i$ in $N_u^1$ |
| $D$ | Degree of a randomly selected node |
| $D_g$ | Degree of a randomly selected generator node |
| $D_g^1$ | Degree of a randomly selected node in $N_g^1$ |
| $D_g^2$ | Degree of a randomly selected node from $N_g^2$, that is itself a neighbor of a node in $V(u) \cap N_u^1$ |

https://doi.org/10.1371/journal.pone.0241790.t001
represents the size of $V(u)$. Moreover, let $X_1^d$ denote a random variable that represents the number of nodes in $V(u) \cap N_1$. Similarly, let $X_2^d$ denote a random variable that represents the number of nodes in $V(u) \cap N_2$, that are themselves neighbors of node $v_i$, located in $V(u) \cap N_1$. Note that $X_2^d$ and $X_2^j$ are independent and identically distributed random variables.

Note that if conditions 3 and 4 of Assumption 1 are satisfied, then

$$X_d = \left( \sum_{i=1}^{x_1^d} X_2^i \right) + X_1^d + 1 \quad (2)$$

The first term in Eq (2) characterizes the number of nodes in $N_2$, which depends on the realization of $X_1^d$. The term $X_1^d + 1$ characterizes the number of nodes in $N_1$ plus the generator node. Let $F_{1d}(x) = P[X_1^d = x]$ and $F_{2d}(x) = P[X_2^d = x]$ represent the pmfs of $X_1^d$ and $X_2^d$. Moreover, let $k_m$ represent the minimum degree of all nodes. The following theorem characterizes $F_{1d}(x)$.

**Theorem 1.** The pmf of $X_1^d$ is given by

$$F_{1d}(x) = \sum_{i=0}^{x_1^d} P[R_d^i = i] P_1(x, d, P[W_d^i = 1]) \quad (3)$$

where

$$P[R_d^i = x] = P_1(x, d, 1 - p) \quad (4)$$

$$P[Z_d^i = x] = \sum_{d_i = k_m}^{n} P[D_d^i = d_i] P_1(x, d_i - 1, p) \quad (5)$$
\[ P[W_d = x] = \sum_{i=0}^{n} P[Z_d = i] P_2 \left( x, \frac{1}{1+i} \right) \]  

(6)

\[ P_1(x, d, q) = \frac{d}{x} q^{d-x} \]  

(7)

\[ P_2(x, q) = q^{1-x} \]  

(8)

Note that \( F_{1,d}(x) \) can be obtained if \( p \) and the pmf of \( D_{1} \) are known. The proof of Theorem 1 and all other theorems can be found in appendices. Next, Theorem 2 characterizes \( F_{2,d}(x) \).

**Theorem 2.** The pmf of \( X_d^2 \) is given by

\[ F_{2,d}(x) = \sum_{i=0}^{n} P[Y_d^2 = i] P_1(x, i, P[W_d^2 = 1]) \]  

(9)

where

\[ P[R_d^2 = x] = \frac{\sum_{d_j = k}^{u} P[D_d^2 = d_j] P_1(x, d_j - 1, 1 - p) \left( \frac{1}{d_j - i} \right)}{\sum_{k=0}^{n} \sum_{d_j = k+1}^{a} P[D_d = d_j] P_1(k, d_j - 1, 1 - p) \left( \frac{1}{d_j - k} \right)} \]  

(10)

\[ P[G_d^2 = x] = \sum_{d_j = k}^{u} P[D_d = d_j] P_1(x, d_j - 1, 1) \]  

(11)

\[ P[Y_d^2 = x] = \sum_{i=0}^{n} P[R_d^2 = i] P_1(x, i, P[G_d^2 = 0]) \]  

(12)

\[ P[D_d^2 = d_j] = \frac{P[D_d^2 = d_j] P_1(0, d_j - 1, 1 - p)}{\sum_{d_j = k}^{u} P[D_d^2 = d_j] P_1(0, d_j - 1, 1 - p)} \]  

(13)

\[ P[Z_d^2 = x] = \sum_{d_j = k}^{u} P[D_d^2 = d_j] P_1 \left( x, d_j - 1, 1 - \sum_{d_j = k}^{u} P[D = d_j] d_j P_1(0, d_j - 1, 1) \right) \]  

(14)

\[ P[W_d^2 = x] = \sum_{i=0}^{n} P[Z_d^2 = i] P_2 \left( x, \frac{1}{1+i} \right) \]  

(15)

Similarly to \( F_{1,d}(x) \), note that \( F_{2,d}(x) \) can be derived, if in addition to \( p \) and the pmf of \( D_{1} \), the pmf of \( D_{2} \) is known. Let \( F_d(x) = P[X_d = x] \) represent the pmf of \( X_d \). The following result characterizes \( F_d(x) \).

**Theorem 3.** The pmf of \( X_d \) is given by

\[ F_d(x) = \sum_{i=0}^{d} F_{1,d}(i) F_{2,d}(x - i - 1) \]  

(16)
where

\[
F^0_{2d}(x) = \begin{cases} 
1 & x = 0 \\
0 & \text{otherwise}
\end{cases}
\quad (17)
\]

\[
F^i_{2d}(x) = F_{2d}(x) 
\]

\[
F_{2d}(x) = \sum_{m=-\infty}^{\infty} F^{m-i}_{2d}(x-m) F_{2d}(m) 
\quad (19)
\]

Note that if \( F_{2d} \) is known, then \( F^i_{2d} \) can be obtained recursively. Let \( F(x) = P[X = x] \) represent the pmf of the sizes of the Voronoi cells in the case where generator nodes (events) are uniformly distributed. Note that

\[
F(x) = \sum_{d=k_m}^{n} P[D_d = d] F_d(x) 
\quad (20)
\]

Based on Theorems 1-3, we can now compute \( F(x) \) using the following algorithm.

**Algorithm 1 Computing the theoretical pmf of \( X \).**

**Input:** Pmfs of \( D \) and \( D' \), and \( p \).

**Output:** \( F(x) \)

1: \( k_m \leftarrow \) minimum degree of all nodes in \( G \)

2: **for** \( d \leftarrow k_m \) to \( n \) **do**

3: Compute the pmf of \( R'_d, Z'_d, W'_d, \) and \( F_{1d} \) (using Eqs (3)-(6))

4: Approximate the pmf of \( D'_d \) (using Eq (1))

5: Compute the pmf of \( R'_d, G'_d, Y'_d, D'_d, Z'_d, W'_d, \) and \( F_{2d} \) (using Eqs (9)-(15))

6: Compute \( F^0_{2d} \) and \( F^i_{2d} \) (using Eqs (17) and (18))

7: \( F_d(x) = F_{1d}(0) F^0_{2d}(x-1) + F_{1d}(1) F^1_{2d}(x-2) \)

8: **for** \( j \leftarrow 2 \) to \( d \) **do**

9: Compute \( F_{1d} \) (using Eq (19))

10: \( F_d(x) \leftarrow F_d(x) + F_{1d}(j) F^j_{2d}(x-j-1) \)

11: **end for**

12: **end for**

13: Compute \( F \) (using Eq (20))

14: **return** \( F(x) \)

Algorithm 1 enables us to define the following criterion for determining whether the distribution of events in a network obeys a non-uniform distribution.

**Criterion 1 (Hotspot criterion).** Let \( F \) represent the pmf of the sizes of Voronoi cells when generator nodes are uniformly distributed and \( F_e \) the empirical pmf of the sizes of Voronoi cells for a given network.

1. For regular or Poisson networks, there is event concentration if

\[
\chi^2(F_e, F) \geq c(x) 
\quad (21)
\]

where \( c > 0 \) is a threshold (which determines the significance level \( \alpha \) of the Chi Square distribution \( \chi^2 \)).

2. For power law networks, there is event concentration if

\[
\Psi(F_e, F) = \frac{\bar{n}(F) - \bar{n}(F_e)}{\bar{n}(F) - 1} \geq \beta 
\quad (22)
\]
where $\beta > 0$ represents a threshold, and

$$\bar{n}(F) = 4 \left( \frac{\sum_{i=1}^{Q_1(F)-1} F(i)}{Q_1(F)} \right) + \left( \frac{1}{4} - \frac{\sum_{i=1}^{Q_1(F)-1} F(i)}{Q_1(F)} \right) Q_1(F)$$

represents the average size of the cells in the first quartile ($Q_1(F)$) of $F$.

Note that the criterion for identifying hotspots depends on the distribution of the sizes of the Voronoi cells, which in turn depends on the degree distribution of the network. The criterion compares the output distribution of Algorithm 1 with the empirical distributions of the sizes of Voronoi cells. Deviations from $F$ indicate the amount of concentration of events on the network. For regular and Poisson networks, deviations are measured using the $\chi^2$ test. For power law networks, deviations are measured based on the average size of the cells in the first quartiles of $F$ and $F_c$.

### Empirical networks

#### Chicago street network

We use data between January 1 and December 31, 2017, to evaluate the formation of assault hotspots on the street network of the city of Chicago [13]. The street network considers expressways, collectors, and arterials. It has 2902 edges, which represent streets, and 1650 nodes, which represent street intersections. An assault is represented as an event and associated to the nearest intersection. We first consider only handgun assaults and then widen the analysis for all types of assaults reported in the 12 months.

In particular, we evaluate the dispersion of assaults over time based on Criterion 1. To identify the length of the observation period that is required for a stationary, high-concentration outcome to be observed, we consider the following steps.

1. Define an observation period of $t$ days, $t = 7, 14, 21, \ldots$
2. Consider assaults reported within $t$ days after January 1, 2017.
3. Associate each reported assault to the closest street intersection (node) and mark each node with $\epsilon = 1$ or more assaults as a generator node.
4. Apply Criterion 1 to determine if there is a concentration of events (i.e., evaluate whether generator nodes create a significant number of small, adjacent Voronoi cells).
5. Move the observation period by one day and repeat steps 1-4 (i.e, consider the assaults reported within $t$ days after January 2, 2017).
6. Repeat steps 1-5 until the observation period starts on December 31, 2017.
7. For each observation period of length $t$, calculate the percentage of instances (throughout 2017) where the proposed criterion determines a high event concentration (that is, the percentage of instances that the null hypothesis of there being no hotspot formation is rejected).

Since the street network of Chicago resembles a lattice, we need to consider Criterion 1.1, that is, the criterion for detecting hotspot on a regular network. The average local clustering coefficient is 0.07 (a negligible value close to 0), meaning that there are hardly any connections in the network neighborhoods of any node. However, note that for a street network, nodes in $N_1$ share two instead of a single neighbor in $N_1$. As a consequence, condition 4 of Assumption 1 is not satisfied, and the derivation of the random variable $X_d$ (Eq (2)) is no longer a precise
expression. In particular, note that, if nodes in $N_2$ share an additional neighbor, then the first term in Eq (2) will take into account some nodes at level 2 twice. Under such a scenario, the expression for $X_d$ is hard to compute and an important part of our current research efforts. Nonetheless, Eq 2 serves as an approximation for the size of the cell of a randomly selected generator node with degree $d$. Accordingly, Algorithm 1 provides an approximation rather than a precise expression for $F$. Finally, to minimize the number of false positives in detecting hotspots, we use a significance level of $\alpha = 10^{-4}$.

The solid line in Fig 4 represents the percentage of instances that the null hypothesis is rejected for observation periods of different lengths. Note that the minimum period for which the criterion consistently identifies the formation of hotspots throughout 2017 is $t = 21$ days, in which case 100 percent of all 365 evaluations of the null hypothesis are rejected. That is, for an observation period of 21 days (or longer), the proposed criterion suggests that hotspots of handgun assaults are formed over the network. In contrast, the dashed curve in Fig 4 represents the percentage of instances that the null hypothesis is rejected based on the proximity of events in the metric space. In particular, it depicts the outcome of determining whether events are concentrated (i.e., step 4 above) when applying the Hopkins test [14] (instead of Criterion 1). Identifying the formation of hotspots for a Hopkins score below 0.25, the test requires an observation period of more than two months in order to reject the null hypothesis over 90 percent of all instances.

Next, Fig 5 shows the values of the $\chi^2$ test (Eq (21)) for different observation periods. The error bars represent one standard deviation. Note that for observation periods that are longer than two months, the values of $\chi^2$ remain approximately constant, meaning that the amount of event concentration on the network does not change significantly when longer observation periods are considered.

The analysis so far classifies as a generator node any intersection (node) associated on the street network to a single assault (event). That is, $e = 1$. However, it is often of interest to distinguish between intersections where sporadic criminal activities take place and those where criminal activities are comparatively more frequent. We now evaluate the effect of classifying a generator node based on a stronger condition, that is, if at least a particular number of assaults (of any type) is associated to that node within a given observation period.
In particular, we revisit the procedure for determining the length of a period for a stationary, high-concentration outcome to be observed, and consider a varying event threshold \( \varepsilon \geq 1 \) for different observations periods in step 3 above. As before, each reported assault is associated to the closest node in the street network. However, only nodes with a number of assaults of at least \( \varepsilon = t/7 \) are marked as generator nodes for observation periods of length \( t = 7, 14, 21, \ldots \). In other words, an intersection represents a generator node if, on average, more than one assault occurs every 7 days.

Note that defining an event threshold \( \varepsilon \) that varies depending on the length of the observation period represents a stronger condition for classifying generator nodes. Step 3 above marks fewer nodes as generator nodes (compared to the case where \( \varepsilon = 1 \)), since fewer intersections have a persistent high rate of assaults. Nonetheless, the solid line in Fig 6 shows that Criterion
identifies the formation of hotspots for observation periods of any length. The horizontal line indicates that, regardless of the length of the observation period, hotspots on the street network are formed as the outcome of high rates of assaults at intersections which are located relatively close to each other. The dashed line in Fig 6, in contrast, shows that when applying the Hopkins test for the same observation periods, the percentage of instances the null hypothesis is rejected (i.e., hotspot detection) tends to decrease as longer periods are considered. According to the Hopkins test, intersections with high rate of assaults are not close to each other in the metric space. Unlike the proposed approach, the Hopkins test is not a decisive approach to identify a high concentration of events on the network (it now rejects the null hypothesis only about 40-50% of all instances).

Co-purchase network of Amazon

Next, consider a co-purchase network of products from Amazon [15]. After a purchase, users can rate their satisfaction with the product they bought on a scale from 1 (very dissatisfied) to 5 (very satisfied). The co-purchase network consists of nodes that represent users and edges that connect users who bought the same product. The average rating of a user indicates overall user satisfaction. Dissatisfied users are defined as the users whose average rating is below the 10th percentile. Analyzing event concentration on the user co-purchase network enables us to evaluate whether dissatisfied users, who purchase a shared set of products, are concentrated on some parts of the network. The network contains 8444 nodes and 38492 edges.

The average local clustering of the network is 0.68, which implies that the network does not satisfy Assumption 1. Fig 7 assesses the quality of the approximation provided by Algorithm 1. It shows the percentage of simulations where the null hypothesis is accepted based on the Chi Square test for different significance levels \(\alpha\). Note that with \(\alpha = 10^{-4}\) the null hypothesis is accepted more than 90% of all simulations. In other words, for a significance level \(\alpha \leq 10^{-4}\), the simulated distribution is equal to the theoretical approximation for more than the 90% of the runs. Though Assumption 1.3 is not satisfied, the approximation provided by Algorithm 1 is quite good.
Fig 8 shows the complementary cumulative degree distribution. Given the power law that characterizes the tail of the degree distribution, we use Criterion 1.2 to determine whether events are concentrated.

Fig 9 shows the pmf of the distribution of the sizes of Voronoi cells and the theoretical distribution. The proposed criterion determines that dissatisfied users are not concentrated. Indeed, note that the empirical distribution resembles the theoretical distribution for which events are located uniformly at random. If we apply the $\chi^2$ test to both distributions, then the null hypothesis is accepted (for $\alpha = 10^{-4}$), which means that dissatisfied users are uniformly distributed across the network.

Finally, we modify the initial ratings to obtain an artificial concentration of dissatisfied users. To generate these hotspots, we first divide the co-purchase network into communities based on the measure of community modularity. Second, we select two communities such that the total number of members of both communities is approximately 844 ($p = 0.1$). Third, we select products that have been bought by at least two members of the two communities, and assign a rating of 1 to the transactions that involve these products. Finally, we compute the set of dissatisfied users based on the new average rating of each user. Note that dissatisfied users are now arbitrary concentrated across the two selected communities. Criterion 1.2 determines that dissatisfied users are now indeed concentrated. Fig 10 shows the pmf of the sizes of Voronoi cells when dissatisfied users are marked as events and the theoretical pmf from a uniform allocation. As expected, the number of cells of small size increases when events are concentrated. Note that the number of cells of size one in the empirical distribution is approximately four times larger than the number in the theoretical distribution.
Discussion

The proposed framework enables us to derive a summary statistic for measuring event concentration based on Voronoi diagrams. It provides an approximation for the distribution of the sizes of Voronoi cells for regular, Poisson, and power law networks in which events are distributed uniformly at random. When the distribution of events obeys a non-uniform allocation, groups of small, adjacent Voronoi cells indicate subnetworks where events (generator nodes) are highly concentrated (hotspots are formed).

Building on this key property of Voronoi diagrams, the proposed criterion for detecting hotspots enables us to measure concentration across a variety of scenarios in which events are...
distributed over a network. Its applications range from determining whether events such as traffic accidents or fire outbreaks are concentrated in certain parts of a city, to evaluating whether influencers in a topic area (e.g., sports or politics) are gathered together in particular subgraphs of a social network.

Our work illustrates the criterion by analyzing the distribution of assaults on the street network of Chicago at various time scales, and considering various event thresholds $\varepsilon$. We show how the criterion can be used to estimate the smallest observation period for explaining the formation of stationary concentrations of events. Our analysis of the distribution of events over urban structures such as the street network aims to complement traditional approaches that identify clusters on the metric space. We compare the outcome of the proposed criterion to that of detecting hotspots by evaluating the proximity of events in the metric space (using the Hopkins test). Finally, we also measure event concentration in a co-purchase network and show that dissatisfied users are uniformly distributed over the network.

The results presented in this paper should be considered in the light of some limitations. The theoretical framework used to derive the hotspot criterion requires certain assumptions on the topological properties of the network, which are generally only approximately met by empirical networks. In particular, we assume that (i) the average local clustering coefficient is negligible (below 0.1), and (ii) the degree distribution of the nodes with events resembles that of the entire network. Satisfying these assumptions guarantees that Algorithm 1 can compute the pmf of the sizes of the cells for a network with a uniform event distribution. Analyzing the behavior of the proposed framework for networks with high clustering remains an interesting direction for future research.

Appendix

A: Proof of Theorem 1

**Theorem 1.** The pmf of $X_d^i$ is given by

$$F_{i,d}(x) = \sum_{x_i=0}^d P[R_d^i = i]P_i(x, i, P[W_d^i = 1])$$

where

$$P[R_d^i = x] = P_i(x, d, 1 - p)$$

$$P[Z_d^i = x] = \sum_{d_i = k}^n P[D_d^i = d_i]P_i(x, d_i, 1 - p)$$

$$P[W_d^i = x] = \sum_{i=0}^n P[Z_d^i = i] = iP_2\left(x, \frac{1}{1+i}\right)$$

$$P_i(x, d, q) = \binom{d}{x} q^x(1 - q)^{d-x}$$

$$P_2(x, q) = q^x(1 - q)^{1-x}$$
Proof. To define the pmf of $X_d^{j}$, consider both regular and generator nodes in $N_1$. Let $R_d^{j}$ be a random variable that denotes the number of regular nodes in $N_1$. Since $p$ is the probability of randomly selecting a generator node, we know that

$$ R_d^{j} \sim \text{Bin}(d, 1 - p) \quad (24) $$

**Remark 1.** Based on definition 1, if a regular node $v_i$ satisfies that $p(v_i, u) \leq \rho(v_i, u_j)$ for all $u_j \in U$, and $i = |\{u_j \in U \setminus \{u\}: \rho(v_i, u) = \rho(v_i, u_j)\}|$, then the probability that node $v_i$ belongs to $V(u)$ is $\frac{1}{n+1}$. Otherwise, the probability that $v_i$ belongs to $V(u)$ is 0.

Note that a regular node in $N_1$ does not necessarily belongs to $V(u)$. According to remark 1, the probability that a regular node in $N_1$ belongs to $V(u)$ depends on the number of neighboring generator nodes of that node. Let $Z_d^{j}$ be a random variable that represents the number of generator nodes in $N_2$, that are neighbors of a regular node in $N_1$.

**Remark 2.** Note that the distribution of the number of generator neighbors in $N_2$ of a regular node with degree $d$, located in $N_0$, obeys a binomial distribution $\text{Bin}(d_1 - 1, \rho)$.

According to Remark 2, the pmf of $Z_d^{j}$ is a mixture of binomial distributions

$$ P[Z_d^{j} = x] = \sum_{d_1 = k_d}^{n} P[D_d^{j} = d_1] P_{1}(x, d_1 - 1, \rho) \quad (25) $$

where

$$ P_{1}(x, d, q) := \binom{d}{x} q^x (1 - q)^{d-x} $$

represents the pmf of the Binomial distribution.

Let $W_d^{j}$ be a Bernoulli random variable that indicates if a regular node in $N_1$ belongs to $V(u)$. According to Remark 1, note that

$$ P[W_d^{j} = x] = \sum_{i=0}^{n} P[Z_d^{j} = i] P_{2}(x, i, \rho) \quad (26) $$

where

$$ P_{2}(x, q) := q^x (1 - q)^{1-x} $$

represents the pmf of the Bernoulli distribution.

Note that $P[W_d^{j} = 1]$ denotes the probability that a regular node in $N_1$ belongs to $V(u)$. Moreover, if there are a total of $i$ regular nodes in $N_1$, then $\text{Bin}(i, P[W_d^{j} = 1])$ is the distribution of the number of nodes that belong to the cell in $N_1$. The distribution of $X_d^{j}$ obeys

$$ F_{1,i}(x) = \sum_{i=0}^{d} P[R_d^{j} = i] P_{1}(x, i, P[W_d^{j} = 1]) \quad (27) $$

**B: Proof of Theorem 2**

**Theorem 2.** The pmf of $X_d^{j}$ is given by

$$ F_{2,i}(x) = \sum_{i=0}^{n} P[Y_d^{j} = i] P_{1}(x, i, P[W_d^{j} = 1]) \quad (9) $$
where

\[
P[R_d^2 = x] = \frac{\sum_{i=1}^{n} P[D_j = d_j] P_1(x, d_j - 1, 1 - p) \left( \frac{1}{d_j - i} \right)}{\sum_{k=0}^{n} \sum_{j=k+1}^{n} P[D_j = d_j] P_1(k, d_j - 1, 1 - p) \left( \frac{1}{d_j - k} \right)},
\]

(10)

\[
P[G_d = x] \approx \sum_{d_j \neq k_0} P[D = d_j] d_j P_1(x, d_j - 1, p)
\]

(11)

\[
P[Y_d^2 = x] = \sum_{i=0}^{n} P[R_d^2 = i] P_1(x, i, P[G_d^2 = 0])
\]

(12)

\[
P[D_d^2 = d_j] = \frac{P[D_d^2 = d_j] P_1(0, d_j - 1, 1 - p)}{\sum_{d_j \neq k_0} P[D_d^2 = d_j] P_1(0, d_j - 1, p)}
\]

(13)

\[
P[Z_d^2 = x] = \sum_{d_j \neq k_0} P[D_d^2 = d_j] P_1(x, d_j - 1, 1 - \sum_{d_j \neq k_0} P[D = d_j] d_j P_1(0, d_j - 1, p)}
\]

(14)

\[
P[W_d^2 = x] = \sum_{i=0}^{n} P[Z_d^2 = i] P_2\left(x, \frac{1}{i+1}\right)
\]

(15)

**Proof.** We extend the analysis described in the appendix A: Proof of Theorem 1 of N_1 to N_2. Let R_d^2 denote a random variable that represents the number of regular nodes, located in N_2, that are neighbors of a single node in V(u) \cap N_1. Note that a node with degree d_j in N_1 has, with probability P_i(i, d_j - 1, 1 - p), a total of i regular neighbors in N_2. Based on Remark 1, this node belong to V(u) with probability \( \frac{1}{\delta} \). The probability that a node that belongs to V(u), located in N_1, has i regular nodes in N_2 is given by

\[
P[R_d^2 = x] = \frac{\sum_{d_j \neq k_0} P[D_d^2 = d_j] P_1(x, d_j - 1, 1 - p) \left( \frac{1}{d_j - i} \right)}{\sum_{k=0}^{n} \sum_{d_j \neq k+1}^{n} P[D_d^2 = d_j] P_1(k, d_j - 1, 1 - p) \left( \frac{1}{d_j - k} \right)}
\]

(28)

Note that the probability that a regular node, located in N_1 of node u, belongs to V(u) is greater than 0. However, according to Remark 1, this is not true for N_2 when \( \delta \geq 2 \). For instance, if a regular node, located in N_2, has a neighboring generator node, located in N_3, then the probability that the regular node belongs to V(u) is 0. We refer to a regular node that has a probability greater than 0 to belong to V(u), as a *candidate node*. In other words, a regular node \( v \) is a candidate node if for all \( u' \in U: \rho(v, u) \leq \rho(v, u') \). Note that each regular node, located in N_1, is a candidate node. Fig 11 highlights the candidates nodes in green.

Consider a randomly selected regular node \( v \), located in N_2, that is a neighbor of a single node in V(u) \cap N_1. Let G_d^2 denote a random variable that represents the number of neighboring generator nodes of v, located in N_3. According to Remark 2 and using the approximation
given in Eq (1), note that

\[
P(G_2 \neq 0) = \sum_{d_i = k_m}^{d_i} p(D_i = d_i) P_1(x, d_i - 1, p) \approx \sum_{d_i = k_m}^{d_i} \frac{P(D_i = d_i)}{k} P_1(x, d_i - 1, p) \tag{29}
\]

Furthermore, let \(Y_2 \) denote a random variable that represents the number of candidate nodes in \(N_2\) that are neighbors for a single node that belongs to \(V(u) \cap N_1\). A regular node in \(N_2\) is a candidate node if there is not a neighboring generator node in \(N_3\). So the probability of a regular node in \(N_2\) of being a candidate node of \(V(u)\) is \(P(G_2 = 0)\). Then, note that if a node \(v\), located in \(N_1\), has \(i\) neighboring regular nodes, the distribution of the number of neighboring candidate nodes in \(N_2\) for node \(v\) obeys Bin\((i, P(G_2 = 0)\)). That is

\[
P(Y_2 = x) = \sum_{i=0}^{x} P(R_i = i) P_1(x, i, P(G_2 = 0)) \tag{30}
\]

Based on the distribution of candidate nodes in \(N_2\), we now calculate the distribution of the number of nodes that belong to \(V(u) \cap N_2\). Let \(D_2\) denote a random variable that characterizes the degree of a randomly selected candidate node in \(N_2\). According to Remark 2, \(P_1(0, d_i - 1, p)\) is the probability that a node with degree \(d_i\), located in \(N_2\), has no generator nodes as neighbors in \(N_3\). So the probability that a randomly selected candidate node in \(N_2\) has degree \(d_i\) is given
by

\[ P[D_d = d] = \frac{\sum_{d=k_u}^{\infty} P[D_d = d] P_i(0, d_i - 1, p)}{\sum_{d=k_u}^{\infty} P[D_d = d] P_i(0, d_i - 1, p)} \quad (31) \]

Note that a candidate node in \( N_2 \) satisfies that its distance to any generator node is greater or equal to 2. Let \( Z^2_d \) denote a random variable that represents the number of generator nodes, except node \( u \), that are at distance of 2 for a candidate node that is located in \( N_2 \). In other words, \( Z^2_d \) represents the number of Voronoi cells, except \( V(u) \), that can contain a candidate node, located in \( N_2 \) of node \( u \). Fig 12 depicts a candidate node of \( V(u) \) that can be potentially contained in three Voronoi cells including \( V(u) \), i.e., the probability that the candidate node belongs to any of the three cells is \( \frac{1}{3} \). Note that for each regular node in \( N_3 \) that has at least one neighboring generator in \( N_4 \), there is an additional Voronoi cell that can potentially contain the candidate node in \( N_2 \). Based on Eq (1), an approximation of the probability that a node in \( N_3 \) has at least one neighboring generator node in \( N_4 \) is

\[ 1 - \sum_{d=k_u}^{\infty} P[D = d] d k P_i(0, d_i - 1, p) \]

Furthermore, note that the distribution of the number of generator nodes, except node \( u \), that can contain in their cells a candidate node in \( N_2 \) with degree \( d_i \) is given by

\[ \text{Bin} \left( d_i - 1, 1 - \sum_{d=k_u}^{\infty} \frac{P[D = d]}{k} P_i(0, d_i - 1, p) \right) \]

According to Remark 1, the pmf of \( Z^2_d \) is the mixture of binomial distributions

\[ P[Z^2_d = x] = \sum_{d=k_u}^{\infty} P[D_d = d] P_i \left( x, d_i - 1, 1 - \sum_{d=k_u}^{\infty} \frac{P[D = d]}{k} P_i(0, d_i - 1, p) \right) \quad (32) \]

Fig 12. The concept of candidate node in \( N_2 \). Representation of a candidate node of \( V(u) \) in \( N_2 \) that can potentially belong to three different Voronoi cells.

https://doi.org/10.1371/journal.pone.0241790.g012
If $W_d^i$ is a Bernoulli random variable that indicates the probability that a candidate node in $N_2$ belongs to $V(u)$, then

$$P[W_d^i = x] = \sum_{i=0}^{n} P[Z_d^i = i] P_2\left(\frac{x}{i+1}\right)$$  \hspace{1cm} (33)

Based on the pmf of $W_d^i$, the distribution of $X_d^i$ is given by

$$F_{d,i}(x) = \sum_{i=0}^{n} P[Y_d^i = i] P_1(x, i, P[W_d^i = 1])$$  \hspace{1cm} (34)

C: Proof of Theorem 3

**Theorem 3.** The pmf of $X_d$ is given by

$$F_d(x) = \sum_{i=0}^{d} F_{d,i}(i) F_{d,i-1}(x - i - 1)$$  \hspace{1cm} (16)

where

$$F_d^0(x) = \begin{cases} 1 & x = 0 \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (17)

$$F_d^1(x) = F_d(x)$$  \hspace{1cm} (18)

$$F_d^j(x) = \sum_{m=-\infty}^{\infty} F_d^{j-1}(x - m) F_d(m)$$  \hspace{1cm} (19)

**Proof.** To define the pmf of $X_d$, recall that $P[X_d^0 = x] = F_{d,0}(x)$. It can be shown that

$$P[X_d^0 + X_d^1 + \cdots + X_d^{j+1} = x] = \sum_{m=-\infty}^{\infty} F_d^j(x - m) F_d(m)$$  \hspace{1cm} (35)

Let $F_d^j(x) = \sum_{m=-\infty}^{\infty} F_d(x - m) F_d(m)$. The distribution of the sum of multiple instances of a random variable can be generalized as

$$P\left[\sum_{j=0}^{\infty} X_d^j = x\right] = \sum_{m=-\infty}^{\infty} F_d^j(x - m) F_d(m)$$  \hspace{1cm} (36)
where

\[
F_{0d}^0(x) = \begin{cases} 
1 & x = 0 \\ 
0 & \text{otherwise}
\end{cases}
\]  

(37)

\[
F_{1d}^1(x) = F_{2d}(x)
\]

(38)

The first term in Eq (36) indicates the probability that \(i\) nodes in \(N_1\) belong to \(V(u)\). The second term represents the pmf of the sum of \(i\) instances of \(X_2^d\), with a shift of \(i + 1\) to account for the number of nodes in \(N_1\) and for the generator node. Note that if the number of nodes in \(N_1\) that belong to \(V(u)\) is zero, then it is not possible that a node in \(N_2\) belongs to \(V(u)\), according to Eq (37).

### Author Contributions

**Conceptualization:** Juan Campos.

**Data curation:** Juan Campos.

**Formal analysis:** Juan Campos.

**Investigation:** Juan Campos.

**Methodology:** Jorge Finke.

**Supervision:** Jorge Finke.

**Validation:** Jorge Finke.

**Writing – original draft:** Juan Campos.

**Writing – review & editing:** Jorge Finke.

### References

1. Eck JE, Liu L. Contrasting simulated and empirical experiments in crime prevention. Journal of Experimental Criminology. 2008; 4(3):195–213. https://doi.org/10.1007/s11292-008-9059-z

2. Groff ER. Simulation for theory testing and experimentation: An example using routine activity theory and street robbery. Journal of Quantitative Criminology. 2007; 23(2):75–103. https://doi.org/10.1007/s10940-006-9021-z

3. Gordon MB, Nadal JP, Phan D, Semeshenko V. Discrete choices under social influence: Generic properties. Mathematical Models and Methods in Applied Sciences. 2009; 19:1441–1481. https://doi.org/10.1142/S0218202509003887

4. Short MB, Brantingham PJ, Bertozzi AL, Tita GE. Dissipation and displacement of hotspots in reaction-diffusion models of crime. Proceedings of the National Academy of Sciences. 2010; 107(9):3961–3965. https://doi.org/10.1073/pnas.0910921107 PMID: 20176972

5. Pitcher AB. Adding police to a mathematical model of burglary. European Journal of Applied Mathematics. 2010; 21:401–419. https://doi.org/10.1017/S0956792510000112

6. Gonzales AR. Mapping Crime: Understanding Hot Spots; 2005. National Institute of Justice report. Available at: http://discovery.ucl.ac.uk/11291/1/11291.pdf.

7. Chaineys R, Ratcliffe J. GIS and crime mapping. John Wiley & Sons; 2013.

8. Buchin K, Cabello S, Gudmundsson J, Löffler M, Luo J, Rote G, et al. Detecting hotspots in geographic networks. In: Advances in GIScience. Springer; 2009. p. 217–231.

9. Campos J, Finke J. Detecting Hotspots on Networks. In: Cherifi H, Gaito S, Mendes JF, Moro E, Rocha LM, editors. Complex Networks and Their Applications VIII. Cham: Springer International Publishing; 2020. p. 633–644.

10. Meyer W. Analyzing Crime on Street Networks: A Comparison of Network and Euclidean Voronoi Methods. University of Illinois. Urbana, Illinois; 2010.
11. Demiryurek U, Shahabi C. Indexing network Voronoi diagrams. Lecture Notes in Computer Science. 2012; 7238:526–543. https://doi.org/10.1007/978-3-642-29038-1_38
12. Fotouhi B, Rabbat MG. Degree correlation in scale-free graphs. The European Physical Journal B. 2013; 86(12):510. https://doi.org/10.1140/epjb/e2013-40920-6
13. Chicago Police Department (2016). Crimes—2001 to Present. Chicago Open Data Portal https://data.cityofchicago.org/Public-Safety/Crimes-2001/8v97-ynyc.
14. Hopkins Brian and Skellam John Gordon A new method for determining the type of distribution of plant individuals Annals of Botany; 1954; 8:213–227.
15. Bing Lu (2010). Amazon ratings. Koblenz Network Collection http://konect.cc/networks/amazon-ratings/