Detecting Hotspots on Networks

Juan Campos(✉) and Jorge Finke

Pontificia Universidad Javeriana, Cali, Colombia
jccampos@javerianacali.edu.cal

Abstract. Traditional approaches for measuring the concentration of events pay little attention to the effects of topological properties. To overcome this limitation, our work develops a theoretical framework to determine whether events are concentrated on a subset of interconnected nodes. We focus on low-clustered networks with regular, Poisson, and power-law degree distributions.

Keywords: Network models · Hotspot formation · Event concentration

1 Introduction

Defining summary statistics for measuring event concentrations enables us to explain general patterns of dispersion. It is not surprising that numerous approaches have been proposed to characterize the distribution of events over space and time [1–4]. Events may represent activities like crime incidents or traffic accidents. To determine whether there exist areas of high event concentration (hotspots), traditional approaches generally evaluate the Euclidean distance between events [3–6].

In particular, kernel-based techniques identify hotspots based on the number of events that fall inside a given neighborhood [4]. Although a wide class of extensions have been proposed [5,6], they do not make use of topological properties of networks on which events can be characterized. For example, events like vehicle thefts, handgun assaults, and traffic accidents are generally constrained to a planar network (e.g., a street network). For scenarios in which events are associated to nodes, the corresponding notion of hotspot depends on the *geodesic* distance between nodes. As a result, efforts to define summary statistics for measuring event concentrations cannot directly apply kernel-based techniques.

It is important to emphasize that the problem of detecting hotspots is not restricted to planar networks. Consider a co-purchase network where nodes represent users and edges indicate that two users have purchased the same product. Ratings associated to users denote overall user satisfaction and events represent users with a very low rating (e.g., two standard deviations below the mean). In this scenario, a concentration of events represents a subset of dissatisfied users who have purchased the same set of products.

H. Cherifi et al. (Eds.): COMPLEX NETWORKS 2019, SCI 881, pp. 633–644, 2020.
https://doi.org/10.1007/978-3-030-36687-2_53

In general, the challenge of identifying high event concentrations on (planar or non-planar) networks arises across a wide range of disciplines. Yet far too little attention has been paid to developing approaches to identify the concentration of events on networks. An exception is the work in [7], which introduces a technique for finding subnetworks that are connected and contain the maximum number of events on the minimal total path length. The approach in [7] focuses on networks with a regular degree distribution, but does not take into account the effects of other degree distributions on the formation of hotspots.

Our research develops a theoretical framework that characterizes the sizes of the Voronoi cells induced by events to determine event concentration. In other words, the framework allows us to evaluate whether a set of events represent the outcome of a non-uniform stochastic allocation process. It can be applied to determine event concentration on low-clustered networks with three particular degree distributions, namely regular, Poisson, and power-laws.

2 Preliminaries

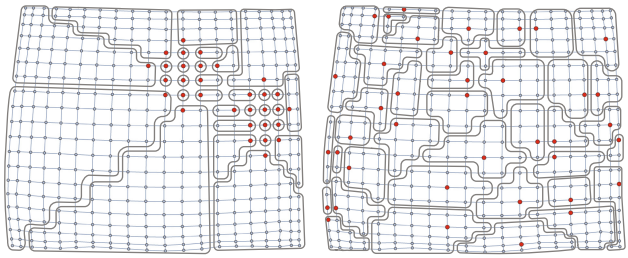
Let $G = (V, E)$ be an undirected network, where $V = \{v_1, \cdots, v_n\}$ represents the set of nodes and $E \subseteq V \times V$ the set of edges. Let $\rho(v_i, v_j)$ denote the geodesic distance between nodes v_i and v_j. Moreover, consider the following definition of a Voronoi diagram [8].

Definition 1. *Let $U = \{u_1, \cdots, u_m\} \subseteq V$ denote a set of generator nodes. The Voronoi diagram of $G = (V, E)$ generated by U is a partition $\{V(u_1), \cdots, V(u_m)\}$ 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 $s' \in \{1, \cdots, m\}$. If $\rho(v_i, u_s) = \rho(v_i, u_{s'})$, then node v_i is assigned to either $V(u_s)$ or $V(u_{s'})$ with equal probability.*

A generator node u_i represents a node that is associated with the occurrence of an event (for example, a crime or accident on the intersection of a street network). Regular nodes represent nodes at which no event occurs and belong to the set $U^c = V - U$. Each element of the Voronoi diagram is called a *cell*. Note that cell $V(u_s)$ contains exactly one generator node, namely u_s. Note also that if G is connected, then every regular node $v_i \in V$ belongs to a cell.

Let $n_s = |V(u_s)| \geq 1$ denote the size of $V(u_s)$. Based on Definition 1 and the distribution of n_s for all $u_s \in U$, that is the distribution of the sizes of the cells, we aim to determine whether events on G are uniformly distributed. Deviations from the uniform distribution will indicate a concentration that results from a non-uniform event allocation.

Consider the regular network in Fig. 1(a), in which a non-uniform allocation yields relative small geodesic distances between generator nodes. Compared to Fig. 1(b), where generator nodes are uniformly distributed, most cells of the Voronoi diagram in Fig. 1(a) contain a small number of regular nodes. Figure 2 shows the probability mass function (pmf) of the sizes of the cells for uniform and non-uniform event allocations.



(a) (b)

Fig. 1. Resulting Voronoi cells when generator nodes (events) are (a) concentrated, and (b) located uniformly at random.

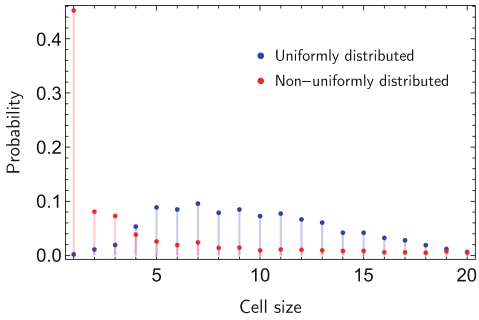


Fig. 2. Probability distributions of the sizes of the Voronoi cells for a network in which events are uniformly distributed (blue) compared to a network with high event concentration (red).

Let D denote a random variable that represents the degree of a randomly selected node of G. Consider a randomly selected node v, with degree $d_v = d$. Let $N_\delta^v = \{v' \in V : \rho(v, v') = \delta\}$ represent the neighborhood of nodes located at a distance δ from node v. Moreover, let D_d^δ denote a random variable that represents the degree of a randomly selected node in N_δ^v.

To derive the pmf of the sizes of the Voronoi cells, consider assumption 1.

Assumption 1. *Suppose the following statements are true:*

1. *The probability that a randomly selected node has degree d is known.*
2. *The probability that a randomly selected neighbor of a node with degree d has degree d' is known.*
3. *The probability that a randomly selected node, located at a distance greater than 1 from a node of degree d, has degree d', can be approximated by the probability that the end node of a randomly selected edge is a node of degree d'.*

Assumptions 1.1 and 1.2 require minimum conditions on the network, namely, that the degree distribution (the pmf of D) and the conditional degree distribution (the pmf of D_d^1) be known. Conditional degree distributions are characterized for various network models and empirical networks [9]. Assumption 1.3

requires that the conditional degree distribution for nodes located at a distance greater than one from a node of degree d does not depend on the degree of that node. Under assumption 1.3, the pmf of D_d^δ for $\delta \geq 2$, can be approximated as

$$P[D_d^\delta = d_i] \approx \frac{nP[D = d_i]\, d_i}{2|E|} = \frac{P[D = d_i]\, d_i}{\bar{k}} \tag{1}$$

where $\bar{k}$ is the average degree of the entire network. The numerator $nP[D = d_i]\, d_i$ represents the total number of end nodes of all edges that are of degree d_i. The denominator $2|E|$ represents the total number of end nodes of the edges of the network.

3 Theoritical Framework

Let D_g denote a random variable that represents the degree of a randomly selected generator node. Moreover, let $p = m/n,\ 0 < p < 1$, denote the proportion of nodes that are generator nodes. Consider a randomly selected generator node, denoted as node u. For convenience, let N_δ denote N_δ^u.

Assumption 2. *The following statements are true:*

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.*
4. *Nodes in N_2 have a single neighbor in N_1.*

Let $d_u = d$ denote the degree of generator node u. Note that the random variable D_d^δ represents the degree of a randomly selected node in $N_\delta = N_\delta^u$. Let X_d denote a random variable that represents the size of $V(u)$. Moreover, let X_d^1 denote a random variable that represents the number of nodes in $V(u) \cap N_1$. Similarly, let X_d^{2i} denote a random variable that represents the number of nodes in $V(u) \cap N_2$, that are neighbors of node v_i, located in $V(u) \cap N_1$. Note that X_d^{2i} and X_d^{2j} are independent and identically distributed random variables.

According to assumptions 2.3 and 2.4, note that

$$X_d = \left(\sum_{i=1}^{X_d^1} X_d^{2i} \right) + X_d^1 + 1 \tag{2}$$

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

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

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

where

$$\begin{aligned}
P[R_d^1 = x] &= P_1(x, d, 1-p), \\
P[Z_d^1 = x] &= \sum_{d_i = k_m}^{n} P[D_d^1 = d_i] P_1\big(x, d_i - 1, p\big) \\
P[W_d^1 = x] &= \sum_{i=0}^{n} P[Z_d^1 = i] P_2\left(x, \frac{1}{1+i}\right), \\
P_1(x, d, q) &= \binom{d}{x} q^x (1-q)^{d-x}, \\
P_2(x, q) &= q^x (1-q)^{1-x}
\end{aligned}$$

The proof of all theorems can be found in the extended version of this paper. Note that $F_{1d}(x)$ can be obtained if p and the pmf of D_d^1 are known. Next, Theorem 2 characterizes $F_{2d}(x)$.

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

$$F_{2d}(x) = \sum_{i=0}^{n} P[Y_d^2 = i] P_1(x, i, P[W_d^2 = 1]) \tag{4}$$

where

$$\begin{aligned}
P[R_d^2 = x] &= \frac{\sum_{d_j=i+1}^{n} P[D_d^1 = d_j] P_1(x, d_j - 1, 1-p) \left(\frac{1}{d_j - i}\right)}{\sum_{k=0}^{n} \sum_{d_j=k+1}^{n} P[D_d^1 = d_j] P_1(k, d_j - 1, 1-p) \left(\frac{1}{d_j - k}\right)}, \\
P[G_d^2 = x] &\approx \sum_{d_i = k_m}^{n} \frac{P[D = d_i]\, d_i}{\bar{k}} P_1(x, d_i - 1, p), \\
P[Y_d^2 = x] &= \sum_{i=0}^{n} P[R_d^2 = i] P_1(x, i, P[G_d^2 = 0]), \\
P[\bar{D}_d^2 = d_i] &= \frac{P[D_d^2 = d_i] P_1(0, d_i - 1, p)}{\sum_{d_i = k_m}^{n} P[D_d^2 = d_i] P_1(0, d_i - 1, p)}, \\
P[Z_d^2 = x] &= \sum_{d_i = k_m}^{n} P[\bar{D}_d^2 = d_i] P_1\left(x, d_i - 1, 1 - \sum_{d_i = k_m}^{n} \frac{P[D = d_i]\, d_i}{\bar{k}} P_1(0, d_i - 1, p)\right), \\
P[W_d^2 = x] &= \sum_{i=0}^{n} P[Z_d^2 = i] P_2\left(x, \frac{1}{i+1}\right)
\end{aligned} \tag{5}$$

Let $F_d(x) = P[X_d = x]$ represent the pmf of X_d and note that $F_{2d}(x)$ can be obtained if p and the pmfs of D_d^1 and D_2^2 are known. The following result characterizes $F_d(x)$.

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

$$F_d(x) = \sum_{i=0}^{d} F_{1d}(i) F_{2d}^i(x - i - 1) \tag{6}$$

where

$$F_{2d}^0(x) = \begin{cases} 1 & x = 0 \\ 0 & otherwise \end{cases} \tag{7}$$

$$F_{2d}^1(x) = F_{2d}(x) \tag{8}$$

$$F_{2d}^i(x) = \sum_{m=-\infty}^{\infty} F_{2d}^{i-1}(x - m) F_{2d}(m) \tag{9}$$

Note that if F_{2d} is known, then F_{2d}^i can be obtained recursively. According to Theorem 3, we can derive an expression for $F_d(x)$ using Eqs. 3 and 4. To define the pmf of the sizes of the Voronoi cells, let X denote a random variable that represents the size of the cell of any randomly selected generator node. Moreover, let $F(x) = P[X = x]$ represent the pmf of X. Note that

$$F(x) = \sum_{d=k_m}^{n} P[D_g = d] F_d(x) \tag{10}$$

According to Eq. 10, if D_g and $F_d(x)$ are known, we can derive $F(x)$. Based on Theorems 1, 2, and 3, Algorithm 1 yields the pmf of the sizes of the Voronoi cells when generator nodes (events) are uniformly distributed. The inputs of Algorithm 1 are the proportion of generator nodes, the degree distribution, and the conditional degree distribution of the network.

Figure 3(a) shows compares the cumulative distribution function (cdf) obtained from simulations (dots) against the theoretical distribution (dashed line) for a regular network. Note that the theoretical distribution matches the distribution of the cell sizes of the simulated network. Figure 3(b) illustrates the pmf of the theoretical distribution.

Furthermore, Fig. 4(a) shows compares the cdf from simulations (dots) against the theoretical cdf (dashed line) for a power-law network. Note that some dots do not remain on the dashed line, but are closely located. Figure 4(b) illustrates the pmf of the theoretical distribution.

We measure the similarity between the simulated and the theoretical pmfs using the Pearson's Chi Square test. The test quantifies the likelihood that data obtained from simulations is the result from the theoretical distribution obtained through Algorithm 1. Let $F_s(i)$ represent the proportion of cells of size i, and

Algorithm 1. Computing the theoretical pmf of X.

Input: Pmfs of D and D_1^d, and p.
Output: $F(x)$

```
k_m ← minimum degree of all nodes in G
for d ← k_m to n do
    Compute the pmf of R_d^1, Z_d^1, W_d^1, and F_1d (using eqs. 3 - ??)
    Approximate the pmf of D_d^2 (using eq. 1)
    Compute the pmf of R_d^1, G_d^2, Y_d^2, D̄_d^2, Z_d^2, W_d^2, and F_2d (using eqs. 4 - 5)
    Compute F_2d^0 and F_2d^1 (using eqs. 7 and 8)
    F_d(x) = F_1d(0)F_2d^0(x − 1) + F_1d(1)F_2d^1(x − 2)
    for j ← 2 to d do
        Compute F_2d^j (using eq. 9)
        F_d(x) ← F_d(x) + F_1d(j)F_2d^j(x − j − 1)
    end for
end for
Compute F (using eq. 10)
return F(x)
```

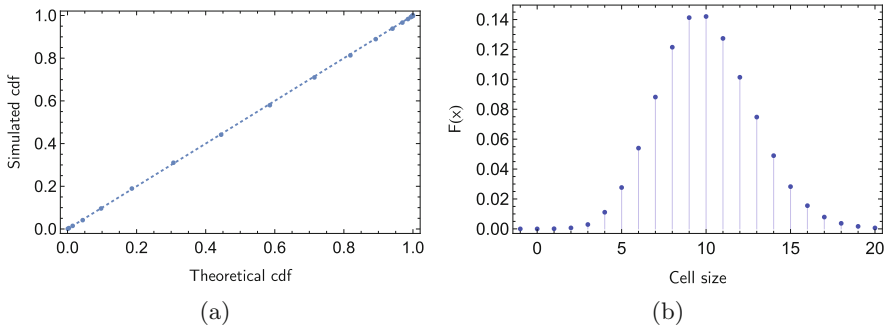
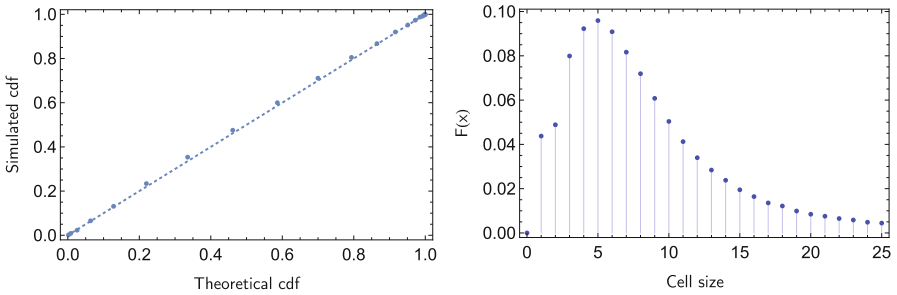


Fig. 3. Theoretical distribution of the sizes of the Voronoi cells, obtained from Algorithm 1 for a regular network with $n = 20000$, $d = 15$ and $p = 0.1$. (a) Probability plot of simulated distribution that results from 100 simulation runs. (b) Plot of the theoretical distribution $F(x)$.

N_s the set of the cell sizes obtained through each simulation. For a network with m generator nodes, the test-statistic value is defined by

$$\chi^2(F_s, F) = m \sum_{i \in N_s} \frac{(F_s(i) - F(i))^2}{F(i)} \tag{11}$$

Figure 5 shows the percentage of simulations where the null hypothesis is accepted for different significance levels α. For regular and Poison networks, a significance level of $\alpha = 0.05$ is sufficient to validate that the simulated distributions resemble the theoretical distribution for more than 95% of all runs. However, power-law networks require a level of significance $\alpha < 10^{-4}$ to validate the null hypothesis for more than the 90% of all runs. Next, we introduce a met-



(a) (b)

Fig. 4. Theoretical distribution of the sizes of the Voronoi cells, obtained from Algorithm 1 for a power-law network with $n = 20000$, and $p = 0.12$. (a) Probability plot of the simulated distribution that results from 100 simulation runs. (b) Plot of the theoretical $F(x)$.

ric which uses theoretical distribution of the sizes of the Voronoi cell to define a formal criterion to determine the concentration of events on a network. In particular, we propose a criterion that uses the theoretical distribution derived in this section to determine whether events are concentrated on a network.

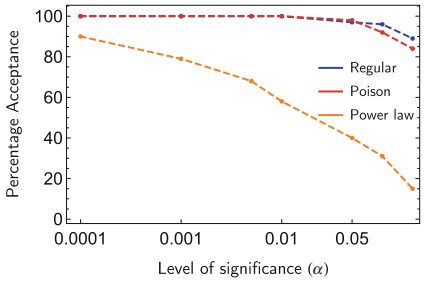


Fig. 5. Percentage of null hypothesis acceptance when generator nodes are located uniformly at random for 100 simulation runs.

4 Event Concentration Criterion

Consider a network G that satisfies assumption 1. Let $F_e(x)$ represent pmf of the sizes of Voronoi cells when events are marked as generator nodes. Moreover, let $F(x)$ be the pmf obtained through Algorithm 1.

Criterion 1. (Hotspots criterion)

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

$$\chi^2(F_e, F) \geq c(\alpha) \tag{12}$$

where c is a threshold value of the Chi Square distribution, which depends on the significance level α.

2. *For power-law networks, there is event concentration if, for $0 < p < 1$,*

$$\Psi(Fe, F) \geq \beta \tag{13}$$

where

$$\Psi(Fe, F) = \frac{\bar{n}(F) - \bar{n}(F_e)}{\bar{n}(F) - 1}, \tag{14}$$

$$\bar{n}(F) = 4\left(\left(\sum_{i=1}^{Q_1(F)-1} F(i)i\right) + \left(0.25 - \sum_{i=1}^{Q_1(F)-1} F(i)\right) Q_1(F)\right) \tag{15}$$

Note that $\bar{n}(F)$ represents the average size of the cells in the first quartile, Q_1, 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. Moreover, note that Algorithm 1 returns F, which we compare 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 χ^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_e.

Next, to evaluate the performance of the proposed criterion, we generate artificial hotspots with different levels of concentration and a constant number of events $m = 2000$. For the method used to generate the artificial hotspots the lower the number of hotspots, the stronger the concentration of the events. Figure 6(a) illustrates Eq. 12 of the criterion for different numbers of hotspots on the regular network. The error bars represent the standard deviations of the χ^2 value for 100 simulations. Figure 6(b) shows the percentage of simulations for which the null hypothesis is rejected, meaning that a hotspot formation is identified. Note that the criterion suggests that if the number of hotspots is greater than 500, then for most cases there is no hotspot detection. According to Fig. 6(b) the null hypothesis is rejected for every simulation and any level of significance α when $h \leq 200$. In other words, the criterion is able to identify the presence of hotspots, when fewer, stronger hotspots emerge (at most $h = 200$).

Next, Fig. 7(a) illustrates Eq. 12 for different numbers of hotspots on the Poisson network. Figure 7(b) shows the percentage of cases where the null hypothesis is rejected. Note that criterion 1.1 identifies the formation of hotspots for every simulation when $h \leq 100$. Compared to regular networks, this suggests that identifying hotspots on a Poisson network requires a stronger concentration of events than for a regular network.

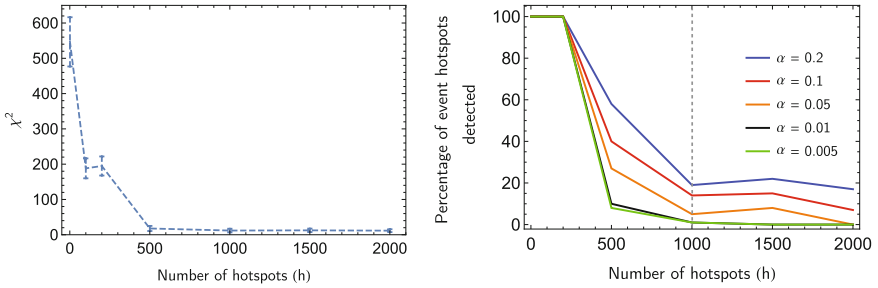


Fig. 6. (a) Value of χ^2 for varying number of hotspots for regular networks ($k = 15$ for 100 simulations). (b) Percentage of rejected null hypothesis for varying number of hotspots.

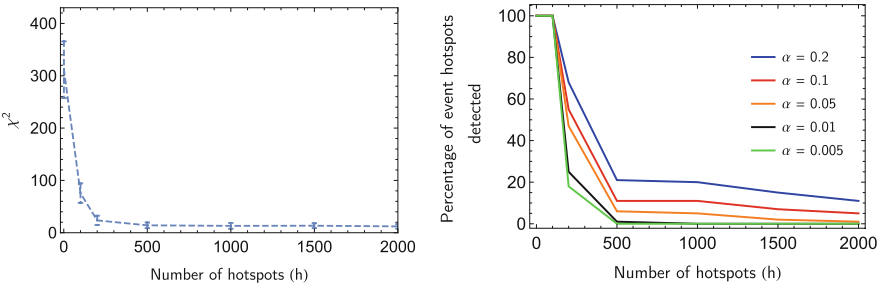


Fig. 7. (a) Value of χ^2 for different number of hotspots for Poisson networks ($q = 0.0015$ for 100 simulations). (b) Percentage of rejected null hypothesis for varying number of hotspots.

According to Figs. 6 and 7, if $\alpha \leq 0.05$, the criterion identifies no concentration of events for all cases when events are located uniformly at random ($h = 2000$), for regular and Poisson networks. Indeed, note that for both networks if $\alpha = 0.01$, then the criterion determines that there is no event concentration for $h \geq 1000$. Furthermore, the criterion identifies event concentration for $h < 500$.

Figure 8(a) illustrates Eq. 14 for different numbers of hotspots. Figure 8(b) shows the performance of the criterion for varying values of β for the power-law network. Note that if $\beta = 0.75$, then the criterion does not identify hotspot formation when $h \geq 1000$. However, if the values of β increases, then the criterion does not identify hotspots for very small values of h.

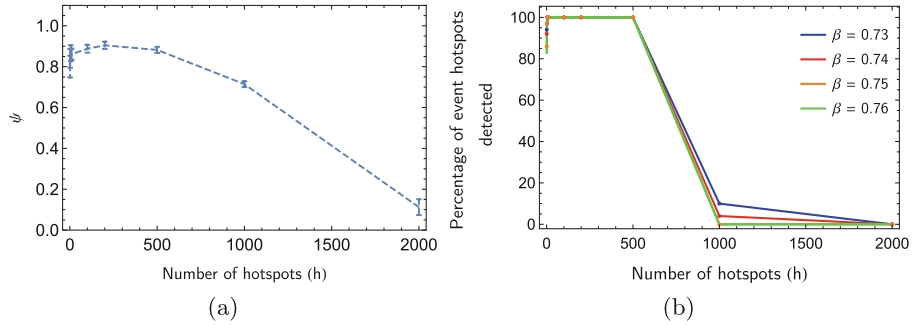


Fig. 8. (a) Eq. 14 for varying number of hotspots for the power-law network. (b) Percentage of instances the criterion determines that there exists event concentration for varying number of hotspots.

Let the optimal β be the minimum value such that criterion 1.2 determines that there is no event concentration when each hotspots is formed by 1 or 2 generator nodes. Note that, according to Fig. 8(b), the optimal β for this particular case is 0.75. Figure 9 illustrates the optimal β, obtained through simulations, for varying proportions of generator nodes. Note that the optimal β does not depend on the proportion of generator nodes.

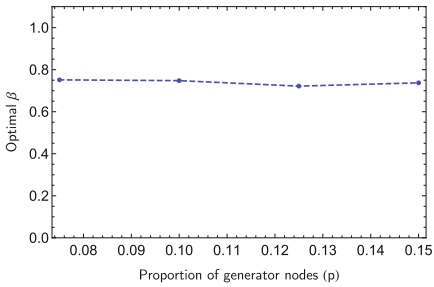


Fig. 9. Optimal β for varying proportion of generator nodes.

5 Conclusions

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 generated by uniformly random events on networks with regular, Poisson, and power-law distributions. Comparing the theoretical cell size distribution with the distribution that results from empirical events, the proposed criterion determines whether there exists event concentration. The criterion quantifies deviations between cell size distributions depending on the degree distribution of the network.

Acknowledgments. This research was supported by the Center of Excellence and Appropriation in Big Data and Data Analytics (CAOBA), founded by the Ministry of Information Technologies and Telecommunications of Colombia (MinTIC) and the Colombian Administrative Department of Science, Technology and Innovation (COLCIENCIAS) under grant no. FP44842-anex46-2015.

References

1. Short, M.B., D'orsogna, M.R., Brantingham, P.J., Tita, G.E.: Measuring and modeling repeat and near-repeat burglary effects. J. Quant. Criminol. **25**(3), 325–339 (2009)
2. Nakaya, T., Yano, K.: Visualising crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics. Trans. GIS **14**(3), 223–239 (2010)
3. Gonzales, A.R.: Mapping crime: Understanding hot spots. National Institute of Justice report (2005). http://discovery.ucl.ac.uk/11291/1/11291.pdf. Accessed 1 Nov 2016
4. Chainey, S., Tompson, L., Uhlig, S.: The utility of hotspot mapping for predicting spatial patterns of crime. Secur. J. **21**(1–2), 4 (2008)
5. Short, M.B., D'orsogna, M.R., Pasour, V.B., Tita, G.E., Brantingham, P.J., Bertozzi, A.L., Chayes, L.B.: A statistical model of criminal behavior. Math. Models Methods Appl. Sci. **18**, 1249–1267 (2008)
6. Short, M.B., Brantingham, P.J., Bertozzi, A.L., Tita, G.E.: Dissipation and displacement of hotspots in reaction-diffusion models of crime. Proc. Natl. Acad. Sci. **107**(9), 3961–3965 (2010)
7. Buchin, K., Cabello, S., Gudmundsson, J., Löffler, M., Luo, J., Rote, G., Silveira, R.I., Speckmann, B., Wolle, T.: Detecting hotspots in geographic networks. In: Advances in GIScience, pp. 217–231. Springer (2009)
8. Demiryurek, U., Shahabi, C.: Indexing network Voronoi diagrams. Lecture Notes in Computer Science, vol. 7238, pp. 526–543 (2012)
9. Fotouhi, B., Rabbat, M.G.: Degree correlation in scale-free graphs. Eur. Phys. J. B **86**(12), 510 (2013)