Crime Detection via Crowdsourcing

Daniel L. Pimentel-Alarcón and Claudia R. Solís-Lemus

University of Wisconsin-Madison

Abstract. In this paper we propose a novel yet simple scheme for criminal detection. Rather than tracking the criminal that committed a particular crime, the police will rank the houses suspected to host criminals according to patterns on citizens' tips. We show that this strategy will provably identify the desired houses under reasonable assumptions. This will aid detectives decide where to focus their efforts. We also give related problems of great interest to the community where the same ideas may be applied with similar results. We complement our theoretical findings with experiments that illustrate the effectiveness of this approach.
Keywords. Statistical signal processing; confidence bounds; detection of criminal patterns; crowdsourcing; anonymous tips.

1 Introduction

When there is a major criminal in a neighborhood (drug dealer, kidnapper, serial killer), the police work can be compared to finding a needle in a haystack. The community wants to help, but the number of calls can be overwhelming and the citizens' noblest intentions to contribute can be translated to countless unsubstantiated clues. More importantly, the police cannot follow up all the tips from the community because of limited resources. But what if instead of treating tips as unrelated data, we group them and analyze them to identify patterns?

Recent years have shown us that the active collaboration of a large community, also known as crowdsourcing, can play a decisive role at solving challenging tasks [1, 2]. Examples include finding a lost boat in thousands of satellite images [3], studying migration patterns of birds [4], searching for anomalous archaeological patterns to locate the lost tomb of Genghis Khan [5], propagating information to bring relief in natural disasters [6], tracking stolen vehicles using social media [7], and aiding the transparency and accountability of the justice system [8].

In this paper we formalize the idea of crowdsourcing criminal detection: using the citizens' tips to rank the houses in a community according to the likelihood that they accommodate a criminal. We show that if reasonable assumptions are met, the strategy will provably succeed at locating houses hosting criminals. We extend the model to incorporate major drawbacks like geographic proximity, personal resentment or prejudice, and we will present other settings where similar strategies may be applied with very promising results. We complement our theoretical findings with experiments that illustrate our approach and show its effectiveness.

Organization of the Paper In Section 2 we formally state the problem and our main results, which we prove in Section 3. In Section 4 we present experiments that support our theory. In Section 5 we give a brief discussion of our findings, along with simple generalizations and other settings where our ideas may be applied.

2 Model and Main Results

Suppose there is a criminal that lives in one of the $n+1$ houses of a city. The goal is to identify $h_\star$, the house that hosts the criminal. The police receives m tips from the citizens, and each tip suggests one house suspected to be $h_\star$. In the end we will select the most suggested house, $\hat{h}$.

Let $\mathcal{H} = \{h_1, h_2, \ldots, h_n, h_\star\}$ denote the set of all houses. Suppose that if a citizen provides a tip, he will independently suggest $h_\star$ with probability $p_\star$, and h_j with probability p_j. We will assume without loss of generality that $p_1 \geq p_2 \geq \cdots \geq p_n$. Intuitively, $p_\star$ represents the accuracy of the citizens' perception and p_1 represents their level of prejudice or other sources of inaccuracy.

Our main result is presented in the following theorem. It essentially states that as long as $p_\star$ (the citizens' accuracy) is slightly larger than p_1 (the level of prejudice), then with high probability the most suggested house will indeed be the one hosting the criminal.

Theorem 1. *Let $\epsilon > 0$ be given and suppose*

$$p_\star \;\geq\; p_1 + \sqrt{\tfrac{2}{m}\log\left(\tfrac{n}{\epsilon}\right)}. \tag{1}$$

Then $\hat{h} = h_\star$ with probability at least $1 - \epsilon$.

The proof of Theorem 1 is given in Section 3. Equivalently, Theorem 1 states that as long as we have enough tips to overcome the gap between $p_\star$ and p_1, we will identify $h_\star$ with high probability. This result is related to survey sampling. For a fixed n, the gap between $p_\star$ and p_1 is $O(1/\sqrt{m})$. A conservative two-sample test for difference in proportions q_1 and q_2 states that one is able to distinguish between the two proportions if their confidence intervals do not overlap. The width of each confidence interval is $O(1/\sqrt{m})$ for m the sample size.

2.1 Geographic dependency

In practice, it is more likely that citizens perceive suspicious activities on houses that they frequently see, e.g., neighboring houses or houses on their way to work. We can model this by weighting the *inherent* probabilities $\{p_1, p_2, \ldots, p_n, p_\star\}$ by the exposure that citizens have to the houses, e.g., by the distances between citizens and houses.

To this end we introduce the matrix $\mathbf{G}$ that encodes the information of the geographic dependency. Essentially, $\mathbf{G}$ will specify the proximity of each citizen to each house, and this will determine the probability that each citizen perceives suspicious activities in each house. More precisely, let $\mathbf{G} \in \mathbb{R}^{m\times(n+1)}$. If citizen i lives in house j, then $\mathbf{G}_{ij} := 0$. Otherwise, $\mathbf{G}_{ij}$ denotes how close citizen i is from house j. Intuitively, if citizen i does not live in house j, then the closest citizen i is to house j, the larger $\mathbf{G}_{ij}$. The setup in the previous section is the particular case where all entries in $\mathbf{G}$ are equal.

Example 1. Consider a street with $n+1$ houses. Suppose there is one citizen living in each house, and that each reported one tip to the police, such that $m = n+1$. Suppose we measure the geographic dependency between citizen i and house h_j using the number of houses between h_i and h_j, such that

$$\mathbf{G} \;=\; n+1 - \begin{bmatrix} n+1 & 1 & 2 & 3 & \\ 1 & n+1 & 1 & 2 & \cdots \\ 2 & 1 & n+1 & 1 & \\ 3 & 2 & 1 & n+1 & \\ & \vdots & & & \ddots \end{bmatrix}.$$

In this case, the closer h_i is to h_j, the more likely it is that citizen i notices suspicious activities in house h_j.

Let $\mathbf{p}$ be the diagonal matrix with diagonal elements taking the values in $\{p_1, p_2, \ldots, p_n, p_\star\}$. Let the $(i,j)^{th}$ entry of $\mathbf{P} := \mathbf{Gp}$ (with normalized rows) denote the probability that citizen i suggests h_j (given that citizen i provided a tip). In particular, we use $\mathbf{P}_{i\star}$ to denote the probability that citizen i suggests $h_\star$. In order to present our next result, let us introduce the set of γ-perceptive citizens, defined as

$$\mathcal{C}_\gamma \;:=\; \{\, i \;:\; \mathbf{P}_{i\star} - \mathbf{P}_{ij} \geq \gamma \;\; \forall\, j \}\,.$$

Intuitively, $\mathcal{C}_\gamma$ is the set of citizens that are at least γ more likely to suggest $h_\star$ than any other house.

The next theorem is a generalization of Theorem 1. It states that if there are enough tips from sufficiently perceptive citizens, then with high probability the most suggested house will indeed be the one hosting the criminal.

Theorem 2. *Let $\epsilon > 0$ be given. For any $k \in \mathbb{N}$ define γ_k as:*

$$\gamma_k \;:=\; \tfrac{m-k}{k} + \sqrt{\tfrac{2}{k}\log\left(\tfrac{n}{\epsilon}\right)}. \tag{2}$$

Assume without loss of generality that each citizen provided one tip. If there is a $k \in \mathbb{N}$ such that the set $\mathcal{C}_{\gamma_k}$ has at least k elements, then $\hat{h} = h_\star$ with probability at least $1-\epsilon$.

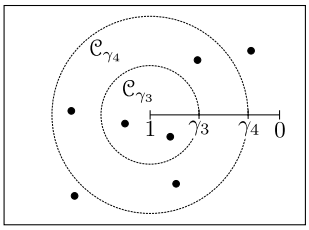

Fig. 1. Theorem 2 asks for a set $\mathcal{C}_{\gamma_k}$ with at least k citizens, such that each of these citizens has a gap between $\mathbf{P}_{i\star}$ and any $\mathbf{P}_{ij}$ at least as large as γ_k. If such set exists, then with high probability we will identify $h_\star$. Notice that γ_k is monotonically decreasing. This implies that $\mathcal{C}_{\gamma_1} \subset \mathcal{C}_{\gamma_2} \subset \cdots$. So the question is: as k grows and γ_k shrinks, will $\mathcal{C}_{\gamma_k}$ grow enough to contain at least k citizens? In this figure, $\mathcal{C}_{\gamma_3}$ only contains 2 citizens (represented with points). It follows that $|\mathcal{C}_{\gamma_3}| = 2 < 3 = k$, and so $\mathcal{C}_{\gamma_3}$ is not large enough to satisfy the conditions of Theorem 2. On the other hand, $\mathcal{C}_{\gamma_4}$ contains 5 citizens. This time $|\mathcal{C}_{\gamma_4}| = 5 > 4 = k$, and so $\mathcal{C}_{\gamma_4}$ satisfies the conditions of Theorem 2. Since there is a set that satisfies these conditions, namely $\mathcal{C}_{\gamma_4}$, we conclude that with high probability we will identify $h_\star$.

The proof of Theorem 2 is given in Section 3. Notice the double dependency on k in Theorem 2. First, k determines the number of perceptive citizens required, that is, the number of citizens in $\mathcal{C}_{\gamma_k}$. And second, it determines how perceptive each of them must be, which is given γ_k. The larger k, the more perceptive citizens are required, but the less perceptive each needs to be.

Also notice that $r := \frac{m-k}{k}$ represents the ratio of non-perceptive citizens versus perceptive citizens (that provided tips). So in words, Theorem 2 states that as long as there is a group $\mathcal{C}_{\gamma_k}$ of k perceptive citizens that are more likely to suggest $h_\star$ over any other house by *a little more* than r, then with high probability we will identify $h_\star$. This *little more* is given by $\sqrt{2/k \log(n/\epsilon)}$. In a nutshell, Theorem 2 requires to have enough citizens that provide tips (at least k) with sufficient accuracy (at least γ_k).

Finally, observe that γ_k is monotonically decreasing. This implies that $\mathcal{C}_{\gamma_{k+1}}$ allows citizens with less perception than $\mathcal{C}_{\gamma_k}$, which in turn implies

$$\mathcal{C}_{\gamma_1} \subset \mathcal{C}_{\gamma_2} \subset \mathcal{C}_{\gamma_3} \subset \cdots .$$

So the question is: as k grows and γ_k shrinks, will $\mathcal{C}_{\gamma_k}$ grow enough to contain at least k citizens? This will depend on $\mathbf{P}$, which in turn depends on $\mathbf{p}$ and $\mathbf{G}$. Fortunately, given $\mathbf{p}$ and $\mathbf{G}$, we can iteratively test whether $\mathcal{C}_{\gamma_k}$ has at least k elements. If so, by Theorem 2 we will identify $h_\star$ with high probability. See Figure 1 to build some intuition.

We point out that Theorem 2 considers the worst-case scenario in which all non-perceptive citizens may even be providing tips collaboratively and adversarially to confuse the police. More about this is discussed in Section 5.

2.2 Tipping Prior

The matrix $\mathbf{P}$ determines how the vote of each citizen *would* be distributed *if* he provided a tip. In this section we add one simple layer to our model to account for the distribution of citizens that provide tips. To this end, observe that

$$\mathbf{P}_{ij} \;=\; \mathsf{P}(\text{citizen } i \text{ suggests } h_j \mid \text{citizen } i \text{ provides a tip})$$

by definition. Letting π_i denote the probability that citizen i provides a tip, it follows that:

$$\mathsf{P}(\text{citizen } i \text{ suggests } h_j) \;=\; \mathbf{P}_{ij}\,\pi_i.$$

It is then clear that the number of citizens that suggest h_j, and hence the outcome of our procedure, will depend on π_i. This probability can be modeled in different ways. For instance, it is reasonable to assume that citizens are more likely to provide a tip if they live near $h_\star$. In this case, we can model π_i as

$$\begin{aligned} &\text{(i)}\ \pi_i \;\propto\; 1/d_{i\star}, \qquad\qquad \text{or} \\ &\text{(ii)}\ \pi_i \;\propto\; \exp\left(-d_{i\star}^2\right). \end{aligned}$$

where $d_{i\star}$ denotes the distance between citizen i and $h_\star$. For example, (ii) corresponds to a gaussian decay in π_i as citizens get far from $h_\star$.

We point out that $\mathbf{G}$ does not capture this information. Without taking into account π_i, this model could yield very poor performance. To see this, suppose that citizen i is so far from $h_\star$, that $\mathbf{P}_{i\star}$ is much smaller than $\mathbf{P}_{ij}$ for some houses h_j neighboring citizen i. But this does not mean that citizen i suspects of any of these houses. In fact, citizen i may not suspect criminal activities in *any* house. In this case, citizen i is unlikely to provide a tip, which equates to π_i being small. But if we ignore π_i, and still ask this citizen to provide a tip, it is very likely (because $\mathbf{P}_{i\star}$ is very small) that he suggests some of his neighboring houses, contaminating the information provided to the police.

3 Proofs

Proof of Theorem 1

Let $N_\star$ and N_j denote the number of suggestions that $h_\star$ and h_j receive. We want to show that with high probability, the criminal lives in $\hat{h}$, the most suggested house. So union bounding over $\mathcal{H}\backslash h_\star$, we have that

$$\mathsf{P}\big(\hat{h} \neq h_\star\big) \;=\; \mathsf{P}\Big(\bigcup_{j=1}^{n}\{N_\star \leq N_j\}\Big) \;\leq\; \sum_{j=1}^{n}\mathsf{P}\big(N_\star \leq N_j\big). \tag{3}$$

Let $Z_j := \frac{1}{m}(N_\star - N_j)$ such that $\mathsf{P}(N_\star \leq N_j) = \mathsf{P}(Z_j \leq 0)$. Letting

$$Z_{ij} \;:=\; \begin{cases} 1 & \text{if } i^{th} \text{ citizen suggested house } h_\star \\ -1 & \text{if } i^{th} \text{ citizen suggested house } h_j \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

it is clear that $Z_j = \frac{1}{m}\sum_{i=1}^{m} Z_{ij}$. Since citizens suggest independently, the Z_{ij}'s are i.i.d. random variables with mean $p_\star - p_j$. Using Hoeffding's inequality [9] we obtain

$$\mathsf{P}\big(Z_j \leq 0\big) \;=\; \mathsf{P}\big(\mathsf{E}[Z_j] - Z_j \geq (p_\star - p_j)\big) \;\leq\; e^{-\frac{m}{2}(p_\star - p_j)^2} \;\leq\; e^{-\frac{m}{2}(p_\star - p_1)^2},$$

where the last inequality follows because $p_1 \geq p_j \;\forall j$ by assumption. Going back to (3), we have that

$$(3) \;=\; \sum_{j=1}^{n} \mathsf{P}\big(Z_j \leq 0\big) \;\leq\; \sum_{j=1}^{n} e^{-\frac{m}{2}(p_\star - p_1)^2} \;<\; n e^{-\frac{m}{2}(p_\star - p_1)^2} \;\leq\; \epsilon,$$

where the last inequality follows by (1). □

Proof of Theorem 2

Let $\mathcal{C}_{\gamma_k}$ be a set satisfying the conditions of Theorem 2. We start as before:

$$\mathsf{P}\big(\hat{h} \neq h_\star\big) \;=\; \mathsf{P}\Big(\bigcup_{j=1}^{n} \{N_\star \leq N_j\}\Big) \;\leq\; \sum_{j=1}^{n} \mathsf{P}\big(N_\star \leq N_j\big). \tag{5}$$

In the worst case scenario, all citizens will most likely suggest the same house (other than $h_\star$), which we will assume without loss of generality to be h_1 (equivalently, $\mathbf{P}_{i1} \geq \mathbf{P}_{ij} \;\forall i, j$). It follows that $\mathsf{P}(N_\star \leq N_j) \;\leq\; \mathsf{P}(N_\star \leq N_1) \;\forall j$, which further implies

$$(5) \;\leq\; n\mathsf{P}\big(N_\star \leq N_1\big) \;=\; n\mathsf{P}\big(Z_1 \leq 0\big), \tag{6}$$

where the last inequality follows by letting $Z_1 := \frac{1}{m}(N_\star - N_1)$. Defining Z_{ij} as in (4), we can write

$$Z_1 \;=\; \frac{1}{m}\sum_{i=1}^{m} Z_{i1} \;=\; \frac{1}{m}\sum_{i \in \mathcal{C}_{\gamma_k}} Z_{i1} \;+\; \frac{1}{m}\sum_{i \notin \mathcal{C}_{\gamma_k}} Z_{i1}.$$

In the worst case scenario, all the non-perceptive citizens will suggest h_1, whence $Z_{i1} = -1$ for every $i \notin \mathcal{C}_{\gamma_k}$. Then

$$Z_1 \;\geq\; \frac{1}{m}\sum_{i \in \mathcal{C}_{\gamma_k}} Z_{i1} \;-\; \frac{m-k}{m},$$

which implies

$$\mathsf{P}\big(Z_1 \leq 0\big) \;\leq\; \mathsf{P}\left(\frac{1}{m}\sum_{i \in \mathcal{C}_{\gamma_k}} Z_{i1} \;\leq\; \frac{m-k}{m}\right) \;=\; \mathsf{P}\left(\frac{1}{k}\sum_{i \in \mathcal{C}_{\gamma_k}} Z_{i1} \;\leq\; \frac{m-k}{k}\right). \tag{7}$$

Letting $Z_1' := \frac{1}{k}\sum_{i\in\mathcal{C}_{\gamma_k}} Z_{i1}$ we obtain

$$(7) \;=\; \mathsf{P}\left(Z_1' \le \frac{m-k}{k}\right) \;=\; \mathsf{P}\left(\mathsf{E}[Z_1'] - Z_1' \ge \mathsf{E}[Z_1'] - \frac{m-k}{k}\right), \qquad (8)$$

and by Hoeffding's inequality [9],

$$(8) \;\le\; e^{-\frac{k}{2}(\mathsf{E}[Z_1'] - \frac{m-k}{k})^2} \;\le\; \frac{\epsilon}{n},$$

where the last inequality follows because $\mathsf{E}[Z_1'] \;=\; \frac{1}{k}\sum_{i\in\mathcal{C}_{\gamma_k}}(\mathbf{P}_{i\star} - \mathbf{P}_{i1})$; by the definition of $\mathcal{C}_{\gamma_k}$, every term of this sum, is at least γ_k, which implies $\mathsf{E}[Z_1']$ is lower bounded by γ_k. We thus conclude that $\mathsf{P}(Z_1 \le 0) \le \frac{\epsilon}{n}$. Substituting this in (6), we obtain the desired result. □

4 Experiments

In this section we present a series of experiments to study the behavior of our detection scheme for different geographic dependency matrices $\mathbf{G}$, which together with the inherent *suspiciousness* level of the houses $\mathbf{p}$, determines the likelihood that citizens perceive suspicious activities. We will test the following cases of $\mathbf{G}$:

(a) **Constant**. This is equivalent to the most basic setup described at the beginning of Section 2, where each citizen suggests each house independently and identically according to $\mathbf{p}$.
(b) **Difference**. Setup described in Example 1, where the geographic dependency is given by the number of houses in between.
(c) **Euclidian**. Same setup as in Example 1, but with the geographic dependency given by the inverse distance, measured in number of houses, i.e.,

$$\mathbf{G} \;=\; \begin{bmatrix} 0 & 1 & 1/2 & 1/3 & \\ 1 & 0 & 1 & 1/2 & \cdots \\ 1/2 & 1 & 0 & 1 & \\ 1/3 & 1/2 & 1 & 0 & \\ & \vdots & & & \ddots \end{bmatrix}.$$

In each trial, we first generate a vector with independent entries selected uniformly at random according to the uniform distribution on $(0,1)^{n+1}$. Next we normalize and sort this vector to obtain $p_\star \ge p_1 \ge p_2 \ge \cdots \ge p_n$. The location of $h_\star$ in the street (and the rest of the houses) is selected uniformly at random. In each trial, m citizens will provide a tip. The citizens that provide tips will be distributed independently over the $n+1$ houses (sample with replacement) according to two tipping priors:

$$\text{(i)}\;\; \pi_i \;:=\; \mathsf{P}(\text{citizen } i \text{ provides a tip}) \;=\; \text{same for every } i,$$

$$\text{(ii)}\;\; \pi_i \;:=\; \mathsf{P}(\text{citizen } i \text{ provides a tip}) \;=\; \exp\left(-\frac{d_{i\star}^2}{400}\right),$$

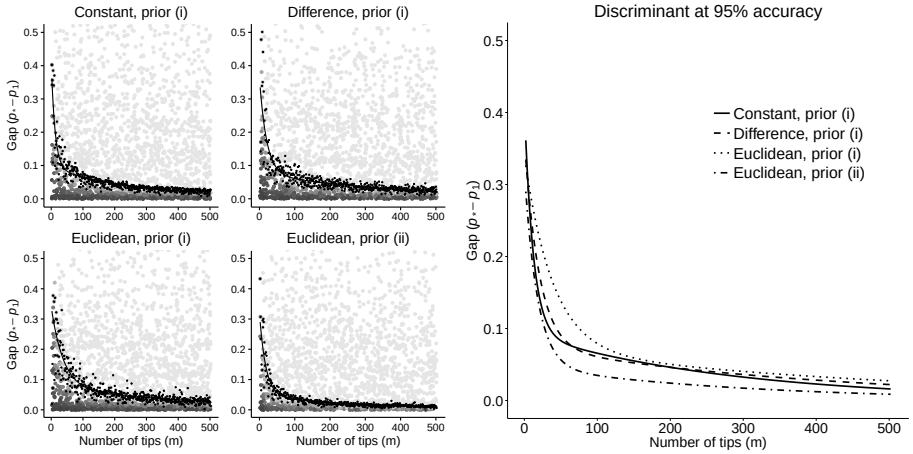

Fig. 2. Left: Phase transition diagram of the success rate at identifying $h_\star$ as a function of the number of tips m and the gap between $p_\star$ and p_1, for four different settings. The gray level at each pair $(m, p_\star - p_1)$ indicates the success rate over $10,000$ replicates: brightest gray represents 100% accuracy; darkest gray represents 0%. Each $(m, p_\star - p_1)$ pair was selected randomly. For each m, all pairs above the black point have at least 95% accuracy. The curve is the best exponential fit to these points. These curves represent the discriminant between at least 95% accuracy (above curve) and less than 95% accuracy (below curve). Intuitively, if we are above the curve, i.e., if we have enough tips, and enough gap between $p_\star$ and p_1, we will likely identify $h_\star$. **Right:** Comparison of the discriminants at 95% accuracy for the four settings in the left. The lower the curve the better, because then fewer tips and gap are required to identify $h_\star$. The Euclidian setting with prior (i) requires more tips and gap, which means identifying $h_\star$ is more difficult. Prior (i) corresponds to the basic model where all citizens are equally likely to provide tips. Prior (ii) corresponds to the more realistic model where citizens near $h_\star$ are more likely to provide tips. Under this model, identifying $h_\star$ requires fewer tips and gap. This can be appreciated by comparing the two Euclidian settings.

where $d_{i\star}$ denotes the distance (measured in number of houses) between h_i and $h_\star$, and 400 represents a variance of roughly 20 houses before the exponential decay. Setting (i) corresponds to the basic model where all citizens are equally likely to provide tips. As discussed in Section 2.2, setting (ii) corresponds to the more realistic scenario where citizens near $h_\star$ are more likely to provide tips.

Each citizen that provides a tip will suggest a house suspected to host criminal activities according to $\mathbf{P}$. Recall that a citizen living in house h_i will suggest house h_j with probability $\mathbf{P}_{ij}$. Since $\mathbf{P} = \mathbf{Gp}$, this probability depends on the house where the citizen lives through the geographic dependency matrix $\mathbf{G}$. We will then select the most suggested house, and we will verify whether it corresponds to $h_\star$. We repeat this experiment $10,000$ replicates for different values of m and $\{p_1, \ldots, p_n, p_\star\}$. The results are summarized in Figure 2.

As predicted by our theory, $h_\star$ can be consistently identified as long as there are enough tips, and there is enough gap between $p_\star$ and p_1. Observe that vis-à-

vis, under prior (i), the Euclidian setting demands more tips and gap than the rest of the settings. This is because the Euclidian matrix $\mathbf{G}$ has a faster decay with distance. We can interpret this as houses being farther apart from one an other. This suggests, in accordance to intuition, that it is easier to find $h_\star$ in denser areas, like highly populated cities, where people are close.

5 Conclusions and Discussion

In this paper we introduce a simple model to identify houses hosting criminals. We prove that under reasonable assumptions, a crowdsourcing strategy will succeed at this task with large probability. Our experiments support our theoretical findings. We now give some simple generalizations to the models described in Section 2, along with other settings where our ideas may be easily extended.

Increasing our Odds. Recall that $p_\star$ and p_1 denote the underlying probabilities that a citizen suggests $h_\star$ and h_1, where h_1 is the most suspicious among the innocent houses. As shown by Theorems 1 and 2, the gap between $p_\star$ and p_1, and the number of citizens that provide tips (m) will determine whether our strategy will work. These quantities can be influenced in our favor through media campaigns to promote participation (to increase m), to encourage citizens to be more aware (to increasing $p_\star$) and to avoid unfounded suggestions, bias or prejudice (to restrict p_1).

Organized Crime. It is also possible that the city has not only one, but several criminals. Moreover, these criminals could be organized and determined to collaborate in an optimal way to avoid detection. In this case, it is in the criminals' best interest to suggest the most suspicious innocent house, h_1. This can be modeled by letting the rows of $\mathbf{P}$ corresponding to criminals take the value 1 in the column corresponding to h_1, and zeros elsewhere. In fact, Theorem 2 is shown assuming that all the citizens not in $\mathcal{C}_{\gamma_k}$ will suggest h_1. Recall that $\mathcal{C}_{\gamma_k}$ denotes the set of perceptive citizens that are at least γ_k more likely to suggest $h_\star$ than any other house.

This implies that Theorem 2 follows regardless of whether the citizens not in $\mathcal{C}_{\gamma_k}$ are criminals or not. We thus conclude that as long as there are enough honest citizens (at least k) with sufficient accuracy (at least γ_k), then with high probability we will find a house hosting a criminal. Hence, we can easily generalize our model to include several criminals' houses. The pattern of identified houses can help detect criminal networks.

Observe that one implicit requirement of Theorem 2 is that the set $\mathcal{C}_{\gamma_k}$ contains at least half of the citizens. This can be seen mathematically because if $k \leq \frac{m}{2}$, then $\frac{m-k}{k} \geq 1$, whence (1) requires that $\gamma_k > 1$, which implies $\mathcal{C}_{\gamma_k} = \emptyset$. In other words, Theorem 2 requires that there are more perceptive citizens than not. This is precisely because Theorem 2 is considering this worst-case adversarial scenario. If there are more organized criminals than honest citizens, then with high probability, h_1 will have more suggestions than $h_\star$.

Detecting Corruption. Of course, none of the ideas discussed above will work if the police force is corrupt. Fortunately, the same ideas can be adapted

to detect patterns of corruption, or equivalently, to find the most honorable policemen. Consider, for an example, the following scenario. Suppose a citizen runs a light and is caught by a policeman. It is the policeman duty to assign a ticket and report it in the system. But if the policeman is corrupt, he will take a bribe and there will be no record of this transaction.

Suppose instead that citizen i runs a light, and is caught by a policeman. An other citizen i' sees that a policeman (who can be identified by the police car) is interacting with the first citizen, and he reports this to the system (anonymously, through a website, a cell phone app, text message, phone call, etc.). Citizen i' does not know the nature of the interaction between the policeman and citizen i, yet he reports that an interaction occurred.

If many citizens report that there was an interaction between a certain policeman, but there is no report of a fine in the system, this would suggest that the policeman took a bribe. If there are many cases suggesting that a particular policeman took bribes, it is likely he did. This would also allow us to identify the most honorable policemen: the ones whose interactions with citizens (reported by citizens) match the fines in the system (reported by the policeman). We can then analyze the hierarchical structure of the corrupt policemen to determine patterns of corruption in higher levels. This will be the case of future study.

References

1. Brabham DC. 2008. Crowdsourcing as a model for problem solving: An introduction and Cases. *Convergence: The International Journal of Research into New Media Technologies*. 14:75–90.
2. Estellés-Arolas E, González-Ladrón-de Guevara F. 2012. Towards an integrated crowdsourcing definition. *Journal of Information Science*. 38:189–200.
3. Doan A, Ramakrishnan R, Halevy AY. 2011. Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*. 54:86.
4. Sullivan B., Wood C., Iliff M., Bonney R., Fink D., Kelling S. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*. 142 (10): 2282 – 2292.
5. Lin AY-M, Huynh A, Lanckriet G, Barrington L. 2014. Crowdsourcing the unknown: The satellite search for Genghis Khan *PLoS ONE*. 9(12): e114046.
6. Goodchild MF, Glennon JA. 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*. 3:231–241.
7. Featherstone C. 2013 Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime *International Conference on Adaptive Science and Technology (ICAST)*. 1–8.
8. Byrne Evans M, O'Hara K., Tiropanis T., Webber C. 2013. Crime applications and social machines: crowdsourcing sensitive data. *In Proceedings of the 22nd International Conference on World Wide Web*. ACM, 891-896.
9. Hoeffding W. 1963. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*. 58 (301):13–30.