Ant-Inspired Density Estimation via Random Walks (Extended Abstact)*

Cameron Musco
cnmusco@mit.edu

Hsin-Hao Su
hsinhao@mit.edu

Nancy Lynch
lynch@csail.mit.edu

Massachusetts Institute of Technology

April 19, 2016

Abstract

Many ant species employ distributed population density estimation in applications ranging from quorum sensing [Pra05], to task allocation [Gor99], to appraisal of enemy colony strength [Ada90]. It has been shown that ants estimate density by tracking encounter rates – the higher the population density, the more often the ants bump into each other [Pra05, GPT93].

We study distributed density estimation from a theoretical perspective. We show that a group of anonymous agents randomly walking on a grid are able to estimate their density d to within a multiplicative factor $1 \pm \epsilon$ with probability $1 - \delta$ in just $\tilde{O}\left(\frac{\log(1/\delta)\log(1/d\epsilon)}{d\epsilon^2}\right)$ steps by measuring their encounter rates with other agents. Despite dependencies inherent in the fact that nearby agents may collide repeatedly (and, worse, cannot recognize when this happens), this bound nearly matches what is required to estimate d by independently sampling grid locations.

From a biological perspective, our work helps shed light on how ants and other social insects can obtain relatively accurate density estimates via encounter rates. From a technical perspective, our analysis requires novel understanding of collision probabilities of multiple random walks using *local mixing properties* of the underlying graph. Our results extend beyond the grid to more general graphs and we discuss applications to biologically-inspired algorithms for social network size estimation, density estimation by robot swarms, and sensor network sampling.

*Full manuscript, including all proofs, available at: http://arxiv.org/abs/1603.02981

1 Introduction

The ability to sense local population density is an important tool used by many ant species. When a colony must relocate to a new nest, scouts search for potential nest sites, assess their quality, and recruit other scouts to high quality locations. A high enough density of scouts at a potential new nest (a *quorum threshold*) triggers those ants to decide on the site and transport the rest of the colony there [Pra05]. When neighboring colonies compete for territory, a high relative density of a colony's ants in a contested area will cause those ants to attack enemies in the area, while a low relative density will cause the colony to retreat [Ada90]. Varying densities of ants successfully performing certain tasks such as foraging or brood care can trigger other ants to switch tasks, maintaining proper worker allocation within in the colony [Gor99, SHG06].

It has been shown that ants estimate density in a distributed manner, by measuring encounter rates [Pra05, GPT93]. As ants randomly walk around an area, if they bump into a larger number of other ants, this indicates a higher population density. By tracking encounters with specific types of ants, e.g. successful foragers or enemies, ants can estimate more specific densities. This strategy allows each ant to obtain an accurate density estimate and requires very little communication – ants must simply detect when they collide and do not need to perform any higher level data aggregation.

1.1 Density Estimation on the Grid

We study distributed density estimation from a theoretical perspective. We model a colony of ants as a set of anonymous agents randomly distributed on a two-dimensional torus graph (which we use in place of a grid to simplify analysis and avoid complications due to boundary behavior).

Computation proceeds in rounds, with each agent stepping in a random direction in each round[1]. A *collision* occurs when two agents reach the same position in the same round and encounter rate is measured as the number of collisions an agent is involved in during a sequence of rounds divided by the number of rounds. Aside from collision detection, the agents have no other means of communication.

The intuition that encounter rate tracks density is clear. It is easy to show that, for a set of randomly walking agents, the *expected* encounter rate measured by each agent is exactly the density d – the number of agents divided by the grid size. However, it is not obvious that encounter rate actually gives a good density estimate – i.e., that it concentrates around its expectation.

Consider agents positioned not on the grid, but on a complete graph. In each round, each agent steps to a uniformly random position and in expectation, the number of other agents they collide with in this step is d. Since each agent chooses its new location uniformly at random in each step, collisions are essentially *independent* between rounds. The agents are effectively taking independent Bernoulli samples with success probability d, and by a standard Chernoff bound, within $O\left(\frac{\log(1/\delta)}{d\epsilon^2}\right)$ rounds obtain a $(1 \pm \epsilon)$ multiplicative approximation to d with probability $1 - \delta$.

On the grid graph, the picture is more complex. If two agents are initially located near each other on the grid, they are more likely to collide via random walking. After a first collision, due to their proximity, they are likely to collide repeatedly in future rounds. The agents cannot recognize repeat collisions since they are anonymous and even if they could, it is unclear that it would help. On average, compared to the complete graph, agents collide with fewer individuals and collide multiple times with those individuals that they do encounter, causing an increase in encounter rate variance and making density estimation more difficult.

Mathematically speaking, on a graph with a *fast mixing time* [Lov93], like the complete graph, each agent's location is only weakly correlated with its previous locations. This ensures that collisions are also weakly correlated between rounds and encounter rate serves as a very accurate estimate of density. The grid graph on the other hand is *slow mixing* – agent positions and hence collisions are highly correlated between rounds. This correlation increases encounter rate variance.

1.2 Our Contributions

Surprisingly, despite this increased variance, encounter rate-based density estimation on the grid is nearly as accurate as on the complete graph. We show that after just $O\left(\frac{\log(1/\delta)\log\log(1/\delta)\log(1/d\epsilon)}{d\epsilon^2}\right)$ rounds, each

[1] Of course, real ant movement is more complex [GPT93, NTD05]. However, our model captures the highly random movement of ants while remaining tractable to theoretical analysis and applicable to ant-inspired random walk-based algorithms

agent's encounter rate is a $(1 \pm \epsilon)$ approximation to d with probability $1 - \delta$.

Technically, to bound accuracy on the grid, we obtain moment bounds on the number of times that two randomly walking agents repeatedly collide over a set of rounds. These bounds also apply to the number of equalizations (returns to starting location) of a single walk. While *expected* random walk hitting times, return times, and collision rates are well understood [Lov93, ES09], higher moment bounds and high probability results are much less common. We hope that beyond the analysis of density estimation, our bounds are of general use in the theoretical study of random walks and random-walk based algorithms.

Our moment bounds show that, while the grid graph is slow mixing, it has sufficiently strong *local mixing* to make random walk-based density estimation accurate. Random walks tend to spread quickly over a local area and not repeatedly cover the same nodes. We discuss applications of our results to a number of random walk-based, biologically-inspired algorithms – for social network size estimation [KLS11], density estimation by robot swarms, and sensor network sampling [AB04, LB07].

2 Density Estimation via Random Walk Collision Rates

We first formally define the density estimation problem.

Density Estimation Problem Consider a two-dimensional torus with A nodes (dimensions $\sqrt{A} \times \sqrt{A}$) populated with $(n+1)$ randomly positioned agents. Define population density as $d \overset{\text{def}}{=} n/A$. Each agent's goal is to estimate d to $(1 \pm \epsilon)$ accuracy with probability $1 - \delta$ for $\epsilon, \delta \in (0,1)$ – i.e., to return an estimate $\tilde{d}$ with $\mathbb{P}\left[\tilde{d} \in [(1-\epsilon)d, (1+\epsilon)d]\right] \geq 1 - \delta$. As a technicality, with $n+1$ agents we define $d = n/A$ instead of $d = (n+1)/A$ for convenience of calculation. In the natural case, when n is large, the distinction is minor.

As discussed, ants solve this problem using an encounter rate-based strategy, formalized in Algorithm 1.

Algorithm 1 Random Walk Encounter Rate-Based Density Estimation

```
input: runtime $t$
  $c := 0$                                                        ▷ Collision counter.
  for $r = 1, ..., t$ do
    $position := position + rand\{(0,1), (0,-1), (1,0), (-1,0)\}$   ▷ Random step (up, down, left, or right).
    $c := c + count(position)$                                      ▷ $count(position)$ = # others at position.
  end for
  return $\tilde{d} = \frac{c}{t}$                                  ▷ Estimate density as encounter rate.
```

Our main result bounds the accuracy of the density estimate returned by Algorithm 1.

Theorem 1 (Random Walk Sampling Accuracy Bound)**.** *After running for t rounds, assuming $t \leq A$, Algorithm 1 returns $\tilde{d}$ such that, for any $\delta > 0$, with probability $\geq 1 - \delta$, $\tilde{d} \in [(1-\epsilon)d, (1+\epsilon)d]$ for $\epsilon = \sqrt{\frac{\log(1/\delta)\log(t)}{td}}$. In other words, for any $\epsilon, \delta \in (0,1)$ if $t = \Theta\left(\frac{\log(1/\delta)\log\log(1/\delta)\log(1/d\epsilon)}{d\epsilon^2}\right)$, $\tilde{d}$ is a $(1 \pm \epsilon)$ multiplicative estimate of d with probability $\geq 1 - \delta$.*

We defer proof to our full writeup [MSL16], briefly discussing some key components of our analysis below. Throughout our analysis, we take the viewpoint of a single agent executing Algorithm 1, referred to as 'agent a'. The first step is to show that the encounter rate $\tilde{d}$ is an unbiased estimator of d:

Lemma 2 (Unbiased Estimator)**.** $\mathbb{E}\,\tilde{d} = d$.

Proof. We can decompose c as the sum of collisions with different agents over different rounds. Specifically, give the n other agents arbitrary ids $1, 2, ..., n$ and let $c_j(r)$ equal 1 if agent a collides with agent j in round r, and 0 otherwise. By linearity of expectation: $\mathbb{E}\,c = \sum_{j=1}^{n}\sum_{r=1}^{t}\mathbb{E}\,c_j(r)$.

Since each agent is initially at a uniform random location and after any number of steps, is still at uniform random location, for all j, r, $\mathbb{E}\,c_j(r) = 1/A$. Thus, $\mathbb{E}\,c = nt/A = dt$ and $\mathbb{E}\,\tilde{d} = \mathbb{E}\,c/t = d$. □

With Lemma 2 in place, the next step is to show that the encounter rate is close to its expectation with high probability and hence provides a good estimate of density.

2.1 Bounding the Effects of Repeat Collisions

Let $c_j = \sum_{r=1}^{t} c_j(r)$ be the total number of collisions with agent j. Due to the initial uniform distribution of the agents, the c_j's are all independent and identically distributed.

Each c_j is the sum of highly correlated random variables – due to the slow mixing of the grid, if two agents collide at round r, they are much more likely to collide in successive rounds. However, by bounding the strength of this correlation, we are able to give strong bounds on the moments of the distribution of each c_j, showing that it is sub-exponential. It follows that $\tilde{d} = \frac{1}{t}\sum_{j=1}^{n} c_j$, is also sub-exponential and hence concentrates strongly around its expectation, the true density d.

We first bound the probability of a re-collision in round $r+m$, assuming a collision in round r:

Lemma 3 (Re-collision Probability Bound)**.** *Consider two agents a_1 and a_2 randomly walking on a two-dimensional torus of dimensions $\sqrt{A} \times \sqrt{A}$. If a_1 and a_2 collide again in round r, for any $m \geq 0$, the probability that a_1 and a_2 collide in round $r+m$ is $\Theta\left(\frac{1}{m+1}\right) + O\left(\frac{1}{A}\right)$.*

Roughly, assuming as in Theorem 1 that $t \leq A$, by Lemma 3, in t rounds, a expects to re-collide with any agent it encounters $\sum_{m=0}^{t-1} \Theta\left(\frac{1}{m+1}\right) = \Theta(\log t)$ times. Our main technical lemma formalizes this intuition, giving a strong moment bound on the distribution of c_j. Intuitively, not only does an agent *expect* to collide at most $O(\log t)$ times with any other agent it encounters, but this bound extends to the higher moments of the collision distribution, and so holds with high probability. In this sense, the grid has strong *local mixing* – random walks spread quickly over a local area and do not cover the same nodes too many times.

Lemma 4. *(Collision Moment Bound) For all $j \in [1,...,n]$, let $\bar{c}_j \overset{\text{def}}{=} c_j - \mathbb{E}\, c_j$. For all $k \geq 2$, assuming $t \leq A$, $\mathbb{E}\left[\bar{c}_j^k\right] = O\left(\frac{t}{A} \cdot k! \log^{k-1} t\right)$.*

2.2 Concentration of Encounter Rate-Based Density Estimate

Armed with the moment bound of Lemma 4 we can show that $\sum_{j=1}^{n} \bar{c}_j$ concentrates strongly about its expectation. Since $\sum_{j=1}^{n} \bar{c}_j$ is just a mean-centered and scaled version of $\tilde{d} = \frac{1}{t}\sum_{j=1}^{n} c_j$, this is enough to prove the accuracy of encounter rate-based density estimation. We first characterize $\sum_{j=1}^{n} \bar{c}_j$ as sub-exponential:

Corollary 5 ($\sum_{j=1}^{n} \bar{c}_j$ is sub-exponential)**.** *Assuming $t \leq A$, $\sum_{j=1}^{n} \bar{c}_j$ is sub-exponential with parameters $b = \Theta(\log t)$ and $\sigma^2 = \Theta(td \log t)$. Specifically, for any λ with $|\lambda| < \frac{1}{b}$ $\mathbb{E}\left[e^{\lambda \sum_{j=1}^{n} \bar{c}_j}\right] \leq e^{\frac{\sigma^2 \lambda^2}{2}}$.*

We finally apply a standard sub-exponential tail bound [Wai15] to prove our main result.

Lemma 6 (Sub-exponential tail bound)**.** *Suppose that X is sub-exponential with parameters (σ^2, b). Then, for any $\Delta \leq \frac{\sigma^2}{b}$, $\mathbb{P}[|X - \mathbb{E}\, X| \geq \Delta] \leq 2e^{-\frac{\Delta^2}{2\sigma^2}}$.*

Proof of Theorem 1. Since $\bar{c}_j$ is just a mean-centered version of c_j, $\sum_{j=1}^{n} \bar{c}_j$ deviates from its mean exactly the same amount as $\sum_{j=1}^{n} c_j$. Further, $\tilde{d}$ is just equal to $\frac{1}{t}\sum_{j=1}^{n} c_j$, so the probability that it falls within an ϵ multiplicative factor of its mean is the same as the probability that $\sum_{j=1}^{n} c_j$ falls within an ϵ multiplicative factor of its mean. By Corollary 5 and Lemma 6:

$$\delta = \mathbb{P}\left[\left|\sum_{j=1}^{n} c_j - \mathbb{E}\left[\sum_{j=1}^{n} c_j\right]\right| \geq \epsilon\, \mathbb{E}\left[\sum_{j=1}^{n} c_j\right]\right] = \mathbb{P}\left[\left|\sum_{j=1}^{n} c_j - td\right| \geq \epsilon td\right] \leq 2e^{\Theta\left(-\frac{\epsilon^2 td}{\log t}\right)}.$$

$\frac{\epsilon^2 td}{\log t} = \Theta\left(\log(1/\delta)\right)$ and so $\epsilon = \Theta\left(\sqrt{\frac{\log(1/\delta)\log t}{td}}\right)$, yielding the theorem. □

3 Algorithmic Applications

Beyond providing a theoretical understanding of how ants estimate density via encounter rates, our analysis has a number of computational implications. In our full writeup, we discuss extensions to higher dimensional tori, hypercubes, and regular expanders. This work helps provide a better understanding of random walk collision rates and highlights the distinction between local and global mixing properties of these graphs.

Additionally, our ant-inspired density estimation algorithm (Algorithm 1), variations on this algorithm, and our analysis techniques can be applied to biologically-inspired techniques for a number of tasks.

3.1 Social Network Size Estimation

Random walk-based density estimation is closely related to work on estimating the size of social networks and other massive graphs [KLS11, KBM12, LL12]. Typically, one does not have access to the full graph (so cannot exactly count the nodes), but can simulate random walks by following links between nodes [MMG$^+$07, GKBM09]. One approach is to run a single random walk and count repeat node visits [LL12, KBM12]. Alternatively, [KLS11] proposes running multiple random walks and counting their collisions. The dominant runtime cost is typically in link queries to the network, and with multiple random walks, this cost can be trivially distributed to multiple servers simulating walks independently.

Walks are first run for a ‘burn-in period’ so that they are distributed by the network’s stable distribution. The walks are then halted, and the number of collisions in this final round are counted. The collision count gives an estimate of the walks’ density, and since the number of walks is known, an estimate of network size.

In our full paper, we show that ant-inspired algorithms can give runtime improvements over this method. After burn-in, instead of halting the walks immediately, we run them for *multiple rounds*, recording encounter rates as in Algorithm 1. This allows the use of fewer walks, decreasing total burn-in cost, and giving faster runtimes when mixing time is relatively slow, as is common in social network graphs [MYK10].

3.2 Distributed Density Estimation by Robot Swarms

Algorithm 1 can be directly applied as a simple and robust density estimation algorithm for robot swarms, and to estimate the frequency of certain properties within the swarm. Let d be the overall population density and d_P be the density of agents with some property P. Let $f_P = d_P/d$ be the relative frequency of P.

Assuming that agents with property P are distributed uniformly in population and that agents can detect this property (through direct communication or some other signal), then they can separately track encounters with these agents. They can compute an estimate $\tilde{d}$ of d and $\tilde{d}_P$ of d_P. By Theorem 1, after running for $t = \Theta\left(\frac{\log(1/\delta)\log\log(1/\delta)\log(1/d\epsilon)}{d_P\epsilon^2}\right)$ steps, with probability $1-2\delta$, $\tilde{d}_P/\tilde{d} \in \left[\left(\frac{1-\epsilon}{1+\epsilon}\right)f_P, \left(\frac{1+\epsilon}{1-\epsilon}\right)f_P\right] = [(1-O(\epsilon))f_P, (1+O(\epsilon))f_P]$ for small ϵ.

In a biological setting, properties may include if an ant has completed a successful foraging trip [Gor99], or if an ant is a nestmate or enemy [Ada90]. In a robotics setting, properties may include whether a robot is part of a certain task group, has completed a task, or has detected an event or environmental property.

3.3 Random Walk-Based Sensor Network Sampling

Finally, we are hopeful that our bounds can be applied to distributed algorithms for sensor network sampling. Random walk-based sensor network sampling [LB07, AB04] is a technique in which a query message (a ‘token’) is initially sent by a base station to some sensor. The token is relayed randomly between sensors, which are connected via a grid network, and its value is updated appropriately at each step to give an answer to the query. This scheme is robust and efficient - it easily adapts to node failures and does not require setting up or storing spanning tree communication structures.

However, if attempting to estimate some quantity, such as the percentage of sensors that have recorded a specific condition, as in density estimation, unless an effort is made to record which sensors have been previously visited, additional variance is added due to repeat sensor visits. Recording previous can be expensive – either the message size must increase or nodes themselves must remember which tokens they have seen. We believe our bounds can be used to show that this is unnecessary – due to strong local mixing, the number of repeat sensor visits will be low, and increased variance due to random walking will be limited.

References

[AB04] Chen Avin and Carlos Brito. Efficient and robust query processing in dynamic environments using random walk techniques. In *Proceedings of the 3rd International Symposium on Information Processing in Sensor Networks*, pages 277–286. ACM, 2004.

[Ada90] Eldridge S Adams. Boundary disputes in the territorial ant *Azteca trigona*: effects of asymmetries in colony size. *Animal Behaviour*, 39(2):321–328, 1990.

[ES09] Robert Elsässer and Thomas Sauerwald. Tight bounds for the cover time of multiple random walks. In *Automata, Languages and Programming*, pages 415–426. Springer, 2009.

[GKBM09] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. A walk in facebook: Uniform sampling of users in online social networks. *arXiv preprint arXiv:0906.0060*, 2009.

[Gor99] Deborah M Gordon. Interaction patterns and task allocation in ant colonies. In *Information Processing in Social Insects*, pages 51–67. Springer, 1999.

[GPT93] Deborah M Gordon, Richard E Paul, and Karen Thorpe. What is the function of encounter patterns in ant colonies? *Animal Behaviour*, 45(6):1083–1100, 1993.

[KBM12] Maciej Kurant, Carter T Butts, and Athina Markopoulou. Graph size estimation. *arXiv preprint arXiv:1210.0460*, 2012.

[KLS11] Liran Katzir, Edo Liberty, and Oren Somekh. Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th International Conference on World Wide Web*, pages 597–606. ACM, 2011.

[LB07] Luisa Lima and Joao Barros. Random walks on sensor networks. In *5th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks and Workshops, 2007*, pages 1–5. IEEE, 2007.

[LL12] Jianguo Lu and Dingding Li. Sampling online social networks by random walk. In *Proceedings of the First ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research*, pages 33–40. ACM, 2012.

[Lov93] László Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdos is Eighty*, 2(1):1–46, 1993.

[MMG$^+$07] Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42. ACM, 2007.

[MSL16] Cameron Musco, Hsin-Hao Su, and Nancy Lynch. Ant-inspired density estimation via random walks. *arXiv preprint arXiv:1603.02981*, 2016.

[MYK10] Abedelaziz Mohaisen, Aaram Yun, and Yongdae Kim. Measuring the mixing time of social graphs. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, pages 383–389. ACM, 2010.

[NTD05] Stamatios C Nicolis, Guy Theraulaz, and Jean-Louis Deneubourg. The effect of aggregates on interaction rate in ant colonies. *Animal Behaviour*, 69(3):535–540, 2005.

[Pra05] Stephen C Pratt. Quorum sensing by encounter rates in the ant *Temnothorax albipennis*. *Behavioral Ecology*, 16(2):488–496, 2005.

[SHG06] Robert J Schafer, Susan Holmes, and Deborah M Gordon. Forager activation and food availability in harvester ants. *Animal Behaviour*, 71(4):815–822, 2006.

[Wai15] Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, draft. http://www.stat.berkeley.edu/~mjwain/stat210b/Chap2_TailBounds_Jan22_2015.pdf, 2015.