Ant-Inspired Density Estimation via Random Walks

[Extended Abstract]*

Cameron Musco
Massachusetts Institute of Technology
32 Vassar Street
Cambridge, MA
cnmusco@mit.edu

Hsin-Hao Su
Massachusetts Institute of Technology
32 Vassar Street
Cambridge, MA
hsinhao@mit.edu

Nancy Lynch
Massachusetts Institute of Technology
32 Vassar Street
Cambridge, MA
lynch@csail.mit.edu

ABSTRACT

Many ant species employ distributed population density estimation in applications ranging from quorum sensing [21], to task allocation [9], to appraisal of enemy colony strength [1]. 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 [21, 10].

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 provides new tools for understanding complex dependencies in the collision probabilities of multiple random walks. We bound the strength of these dependencies using *local mixing properties* of the underlying graph. Our results extend beyond the grid to more general graphs and we discuss applications to social network size estimation, density estimation by robot swarms, and random walked-based sampling of sensor networks.

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 [21]. 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 [1]. Varying densities of ants performing certain tasks such as foraging or brood care can trigger other ants to switch tasks, maintaining proper worker allocation within in the colony [9, 22].

It has been shown that ants estimate density in a distributed manner, via encounter rates [21, 10]. 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 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 grid. Computation proceeds in rounds, with each agent stepping in a random direction in each round. 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 the *expected* encounter rate measured by each agent is exactly the density d – the number of agents divided by the grid size (see Lemma 2). However, it is unclear if encounter rate actually gives a good density estimate – i.e., if 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 $1-\delta$.

*Full version available at http://arxiv.org/abs/1603.02981

PODC'16, July 25 - 28, 2016, Chicago, IL, USA

ISBN 978-1-4503-3964-3/16/07…$15.00

DOI: http://dx.doi.org/10.1145/2933057.2933106

On the grid graph, the picture is significantly more complex. If two agents are initially located near each other, 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 graphs with *fast mixing times* [14], like the complete graph, each agent's location is 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* – agents' 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. $O\left(\frac{\log(1/\delta)\log\log(1/\delta)\log(1/d\epsilon)}{d\epsilon^2}\right)$ rounds suffices for each agent's encounter rate to be a $(1 \pm \epsilon)$ approximation to d with probability $1-\delta$ (see Theorem 1).

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 [14, 5], higher moment bounds and high probability results are much less common. We hope 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. Significant work has focused on showing that random walk sampling is nearly as good as independent sampling for fast mixing expander graphs [7, 4]. We are the first to extend this type of analysis to slowly mixing graphs, showing that strong local mixing is sufficient in many applications.

Beyond the grid, we show how to generate moment bounds from a bound on the probability that two random walks recollide (analogously, that a single random walk equalizes) after a certain number of steps, and apply this technique to d-dimensional grids, regular expanders, and hypercubes. We discuss applications of our results to social network size estimation via random walk [11], obtaining significant improvements over known work for networks with slow global mixing time, but strong local mixing. We also discuss connections to robot swarm density estimation by robot swarms and random walk-based sensor network sampling [3, 13].

2. THEORETICAL MODEL

We consider a two-dimensional torus with A nodes (dimensions $\sqrt{A} \times \sqrt{A}$) populated with identical anonymous agents. We assume that A is large – larger than the area agents traverse over the runtimes of our algorithms. We feel that this torus model successfully captures the dynamics of density estimation on a surface, while avoiding complicating factors of boundary behavior.

Initially each agent is placed independently at a uniform random node in the torus. This placement is important for our bounds – otherwise adversarial positioning could force the agents to walk for the mixing time of the grid $O(A \log A)$ before obtaining good density estimates. We believe the assumption is a reasonable model for the positioning of a colony of active agents looking to perform density estimation, however weakening it would be interesting.

Computation proceeds in discrete, synchronous rounds. Each agent has an ordered pair *position* which it updates in each round with a step chosen uniformly at random from $\{(0,1),(0,-1),(1,0),(-1,0)\}$. Of course, in reality ants do not move via pure random walk – observed encounter rates seem to actually be lower than predicted by a pure random walk model [10, 20]. However, we feel that our model sufficiently captures the highly random movement of ants while remaining tractable to analysis and applicable to ant-inspired random walk-based algorithms (Section 5).

Aside from the ability to move in each round, agents can sense the number of agents other than themselves at their position at the *end of each round*, formally through calling *count*(*position*). We say that two agents *collide in round* r if they have the same position at the end of the round. Outside of collision counting, agents have no means of communication. They are anonymous (cannot uniquely identify each other) and all execute identical density estimation routines.

Density Estimation Problem.

Let $(n+1)$ be the number of agents and define population density as $d \stackrel{\text{def}}{=} n/A$. 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.

3. DENSITY ESTIMATION VIA RANDOM WALK COLLISION RATES

As discussed, the challenge in analyzing random walk-based density estimation arises from increased variance due to repeated collisions of nearby agents. In our full paper [19], we show that, if not restricted to random walking, agents can avoid collision correlations by splitting into 'stationary' and 'mobile' groups and only counting collisions between members of different groups. This allows them to essentially simulate independent sampling of grid locations to estimate density. This method is extremely simple to analyze, however is not 'natural' in a biological sense or useful in the applications of Section 5. Further, independent sampling is unnecessary! Algorithm 1 describes a simple random walk-based approach that gives a nearly matching bound.

3.1 Random Walk Algorithm Analysis

Our main result follows; its proof appears at the end of Section 3 after some preliminary lemmas.

Algorithm 1 Encounter Rate-Based Density Estimation

input: runtime t

$c := 0$
for $r = 1, ..., t$ **do**
  $position := position + rand\{(0,1), (0,-1), (1,0), (-1,0)\}$
  $c := c + count(position)$ ▷ Update collision count.
end for
return $\tilde{d} = \frac{c}{t}$

THEOREM 1 (DENSITY ESTIMATION ACCURACY). *After running for t rounds, for $t \le A$, Algorithm 1 returns $\tilde{d}$ such that, for any $\delta > 0$, with probability $\ge 1 - \delta$, $\tilde{d} \in [(1-\epsilon)d, (1+\epsilon)d]$ 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 $\ge 1 - \delta$.*

Throughout our analysis, we take the viewpoint of a single agent executing Algorithm 1, referred to as 'agent a'. To start, we 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, we now must show that the encounter rate is close to its expectation with high probability and hence provides a good estimate of density.

3.2 Bounding 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 $m \ge 0$, the probability that a_1 and a_2 collide in round $r + m$ is $\Theta\left(\frac{1}{m+1}\right) + O\left(\frac{1}{A}\right)$.*

PROOF. From round r to round $r+m$, a_1 and a_2 take $2m$ random steps in total. Let M_x be the total number of steps they take in the x direction and M_y be the total number in the y direction. $M_x + M_y = 2m$.

We start by computing the probability that the agents collide in round $r + m$ conditioned on the values of M_x and M_y. All steps are chosen independently, so we can consider movement in the x and y directions separately. Specifically, let $\mathcal{C}$ be the event that the a_1 and a_2 collide in round $r+m$, $\mathcal{C}_x$ be the event that they have the same x position, and $\mathcal{C}_y$ be the event that they have the same y position. We have:

$$\begin{aligned}&\mathbb{P}\left[\mathcal{C}|M_x = m_x, M_y = m_y\right]\\&\quad= \mathbb{P}\left[\mathcal{C}_x|M_x = m_x\right]\cdot\mathbb{P}\left[\mathcal{C}_y|M_y = m_y\right].\end{aligned} \tag{1}$$

We first consider $\mathbb{P}\left[\mathcal{C}_x|M_x = m_x\right]$. All bounds will hold symmetrically for the y dimension. We split our analysis into two cases. Let $\mathcal{C}_x^1$ be the event that the two agents have the same x position after round $r + m$ and have identical displacements from their starting locations. Let $\mathcal{C}_x^2$ be the event that the two agents have the same x position after round $r + m$ but *do not* have identical displacements. This requires that the agents 'wrap' around the torus, ending at the same position despite moving different amounts in the x direction. We have $\mathbb{P}[\mathcal{C}_x|M_x = m_x] = \mathbb{P}[\mathcal{C}_x^1|M_x = m_x] + \mathbb{P}[\mathcal{C}_x^2|M_x = m_x]$.

$\mathbb{P}[\mathcal{C}_x^1|M_x = m_x]$ is identical to the probability that a single random walk takes m_x steps and has 0 overall displacement – i.e. takes an equal number of clockwise and counterclockwise steps. This is well known [6] and given by:

$$\mathbb{P}[\mathcal{C}_x^1|M_x = m_x] = \binom{m_x}{m_x/2}\left(\frac{1}{2}\right)^{m_x} = \frac{m_x!}{(\frac{m_x}{2}!)^2}\left(\frac{1}{2}\right)^{m_x}. \tag{2}$$

Above we assume m_x is even – otherwise $\mathcal{C}_x^1$ cannot occur. By Stirling's approximation for any $n > 0$, $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n\left(1 + O\left(\frac{1}{n}\right)\right)$. Plugging this into 2:

$$\begin{aligned}\mathbb{P}[\mathcal{C}_x^1|M_x = m_x] &= \frac{m_x!}{(\frac{m_x}{2}!)^2}\left(\frac{1}{2}\right)^{m_x}\\&= \frac{\sqrt{2\pi m_x}\left(\frac{m_x}{e}\right)^{m_x}\left(1 + O\left(\frac{1}{m_x}\right)\right)}{\pi m_x\left(\frac{m_x/2}{e}\right)^{m_x}\left(1 + O\left(\frac{1}{m_x/2}\right)\right)^2}\cdot\left(\frac{1}{2}\right)^{m_x}\\&= \Theta\left(\frac{1}{\sqrt{m_x + 1}}\right).\end{aligned}$$

(We use $m_x + 1$ instead of m_x in the denominator so that the bound holds in the case when $m_x = 0$.)

$\mathbb{P}\left[\mathcal{C}_x^2|M_x = m_x\right]$ is the probability that two agents have the same x position after round $r + m$ but have different total displacements. It is identical to the probability that a single m_x step random walk has overall displacement $\pm c\sqrt{A}$ for some integer $c \ge 1$ (and so 'wraps around' the torus, ending at its starting location). Roughly, we bound the probability of this event by the probability that the random walk ends at *any* other location on the torus. There are $\sqrt{A}$ such locations, so the probability is bounded by $O\left(\frac{1}{\sqrt{A}}\right)$.

$$\mathbb{P}\left[\mathcal{C}_x^2|M_x = m_x\right] = 2\cdot\left(\frac{1}{2}\right)^{m_x}\cdot\sum_{c=1}^{\left\lfloor\frac{m_x}{\sqrt{A}}\right\rfloor}\binom{m_x}{\frac{m_x - c\sqrt{A}}{2}} \tag{3}$$

where the extra factor of 2 comes from the fact that the displacement may be either clockwise or counterclockwise. (Note that if $\frac{m_x - c\sqrt{A}}{2}$ is not an integer we just define the binomial coefficient to equal 0.)

For $i \in [1, ..., \sqrt{A}-1]$, let $\mathcal{D}_x^i$ be the event that a single random walk is i steps clockwise from its starting location after taking M_x steps. We have:

$$
\begin{aligned}
&\mathbb{P}[\mathcal{D}_x^i | M_x = m_x] \\
&= \left(\frac{1}{2}\right)^{m_x} \cdot \sum_{c=-\left\lfloor \frac{m_x+i}{\sqrt{A}} \right\rfloor}^{\left\lfloor \frac{m_x-i}{\sqrt{A}} \right\rfloor} \binom{m_x}{\frac{m_x+i+c\sqrt{A}}{2}} \\
&\geq \left(\frac{1}{2}\right)^{m_x} \cdot \sum_{c=-\left\lfloor \frac{m_x+i}{\sqrt{A}} \right\rfloor}^{-1} \binom{m_x}{\frac{m_x+i+c\sqrt{A}}{2}} \\
&\geq \left(\frac{1}{2}\right)^{m_x} \cdot \sum_{c=1}^{\left\lfloor \frac{m_x}{\sqrt{A}} \right\rfloor} \binom{m_x}{\frac{m_x+i-c\sqrt{A}}{2}}. \qquad (4)
\end{aligned}
$$

For any $i \in [1, ..., \sqrt{A}-1]$, and any $c \geq 1$, $\frac{m_x+i-c\sqrt{A}}{2}$ is closer to $\frac{m_x}{2}$ than $\frac{m_x-c\sqrt{A}}{2}$ is, so

$$
\binom{m_x}{\frac{m_x+i-c\sqrt{A}}{2}} > \binom{m_x}{\frac{m_x-c\sqrt{A}}{2}} \qquad (5)
$$

as long as $\frac{m_x+i-c\sqrt{A}}{2}$ is an integer. This allows us to lower bound $\mathbb{P}[\mathcal{D}_x^i | M_x = m_x]$ using $\mathbb{P}\left[\mathcal{C}_x^2 | M_x = m_x\right]$. Let $\mathcal{E}_{i,c}$ equal 1 if $\frac{m_x+i-c\sqrt{A}}{2}$ is an integer and 0 otherwise. Since $\mathcal{C}_x^2$ and each $\mathcal{D}_x^i$ are disjoint events:

$$
\mathbb{P}\left[\mathcal{C}_x^2 | M_x = m_x\right] + \sum_{i=1}^{\sqrt{A}-1} \mathbb{P}\left[\mathcal{D}_x^i | M_x = m_x\right] \leq 1
$$

$$
\mathbb{P}\left[\mathcal{C}_x^2 | M_x = m_x\right] + \left(\frac{1}{2}\right)^{m_x} \sum_{i=1}^{\sqrt{A}-1} \sum_{c=1}^{\left\lfloor \frac{m_x}{\sqrt{A}} \right\rfloor} \binom{m_x}{\frac{m_x+i-c\sqrt{A}}{2}} \leq 1
$$

(applying (4))

$$
\mathbb{P}\left[\mathcal{C}_x^2 | M_x = m_x\right] + \left(\frac{1}{2}\right)^{m_x} \sum_{c=1}^{\left\lfloor \frac{m_x}{\sqrt{A}} \right\rfloor} \left(\binom{m_x}{\frac{m_x-c\sqrt{A}}{2}} \sum_{i=1}^{\sqrt{A}-1} \mathcal{E}_{i,c} \right) \leq 1
$$

(by (5) and switching summations)

$$
\mathbb{P}\left[\mathcal{C}_x^2 | M_x = m_x\right] \cdot \Theta(\sqrt{A}) \leq 1.
$$

The last step follows from combining (3) with the fact that $\sum_{i=1}^{\sqrt{A}-1} \mathcal{E}_{i,c} = \Theta\left(\sqrt{A}\right)$ for all c since $\frac{m_x+i-c\sqrt{A}}{2}$ is integral for half the possible $i \in [1, ..., \sqrt{A}-1]$. Rearranging, we have $\mathbb{P}\left[\mathcal{C}_x^2 | M_x = m_x\right] = O\left(\frac{1}{\sqrt{A}}\right)$.

Combining our bounds for $\mathcal{C}_x^1$ and $\mathcal{C}_x^2$, $\mathbb{P}\left[\mathcal{C}_x | M_x = m_x\right] = \Theta\left(\frac{1}{\sqrt{m_x+1}}\right) + O\left(\frac{1}{\sqrt{A}}\right)$. Identical bounds hold for the y direction and by (1) we have:

$$
\begin{aligned}
\mathbb{P}\left[\mathcal{C} | M_x = m_x, M_y = m_y\right] &= \Theta\left(\frac{1}{\sqrt{(m_x+1)(m_y+1)}}\right) \\
&+ O\left(\frac{1}{\sqrt{A(m_x+1)}} + \frac{1}{\sqrt{A(m_y+1)}}\right) + O\left(\frac{1}{A}\right). \qquad (6)
\end{aligned}
$$

Finally, we remove the conditioning on M_x and M_y. Since direction is chosen independently and uniformly at random for each step, $\mathbb{E}\, M_x = \mathbb{E}\, M_y = m$. By a Chernoff bound:

$$
\mathbb{P}[M_x \leq m/2] \leq 2e^{-(1/2)^2 \cdot m/2} = O\left(\frac{1}{m+1}\right).
$$

(Again using $m+1$ instead of m to cover the $m=0$ case). An identical bound holds for M_y, and so, except with probability $O\left(\frac{1}{m+1}\right)$ both are $\geq m/2$. Plugging into (6) gives:

$$
\begin{aligned}
\mathbb{P}\left[\mathcal{C}\right] &= \Theta\left(\frac{1}{m+1}\right) + O\left(\frac{1}{\sqrt{A(m+1)}}\right) + O\left(\frac{1}{A}\right) \\
&= \Theta\left(\frac{1}{m+1}\right) + O\left(\frac{1}{A}\right).
\end{aligned}
$$

□

We note that the techniques of Lemma 3 also apply to bounding the probability that a single random walk returns to its origin (equalizes) after m steps (proof in full paper).

Corollary 4 (Equalization Probability Bound). *Consider agent a_1 randomly walking on a two-dimensional torus of dimensions $\sqrt{A} \times \sqrt{A}$. If a_1 is located at position p after round r, for any even $m \geq 0$, the probability that a_1 is again at position p after 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. By Lemma 2, a expects to be involved in $dt = nt/A$ total collisions. So accounting for re-collisions, it expects to collide with $\Theta\left(\frac{nt}{A \log t}\right)$ unique individuals. This is formalized in Lemma 5 (proof in full paper).

Lemma 5 (First Collision Probability). *Assuming $t \leq A$, for all $j \in [1, ..., n]$, $\mathbb{P}\left[c_j \geq 1\right] = \Theta\left(\frac{t}{A \log t}\right)$.*

We now give our main technical lemma – 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 6. *(Collision Moment Bound) For $j \in [1, ..., n]$, let $\bar{c}_j \stackrel{\text{def}}{=} c_j - \mathbb{E}\, c_j$. For all $k \geq 2$, assuming $t \leq A$, $\mathbb{E}\left[\bar{c}_j^k\right] = O\left(\frac{t}{A} \cdot k! \log^{k-1} t\right)$.*

Proof. We expand $\mathbb{E}[\bar{c}_j^k] = \mathbb{P}[c_j \geq 1] \cdot \mathbb{E}[\bar{c}_j^k | c_j \geq 1] + \mathbb{P}[c_j = 0] \cdot \mathbb{E}[\bar{c}_j^k | c_j = 0]$, and so by Lemma 5:

$$
\mathbb{E}\left[\bar{c}_j^k\right] = O\left(\frac{t}{A \log t} \cdot \mathbb{E}\left[\bar{c}_j^k | c_j \geq 1\right] + \mathbb{E}\left[\bar{c}_j^k | c_j = 0\right]\right).
$$

$\mathbb{E}\left[\bar{c}_j^k | c_j = 0\right] = (\mathbb{E}\, c_j)^k = (t/A)^k \leq \frac{t}{A} k! \log^{k-1} t$ for all $k \geq 2$. Further, $\mathbb{E}\left[\bar{c}_j^k | c_j \geq 1\right] \leq \mathbb{E}\left[c_j^k | c_j \geq 1\right]$, since $\mathbb{E}\, c_j = \frac{t}{A} \leq 1$. So to prove the lemma, it just remains to show that $\mathbb{E}\left[c_j^k | c_j \geq 1\right] = O\left(k! \log^k t\right)$.

Conditioning on $c_j \geq 1$, we know the agents have an initial collision in some round $t' \leq t$. We split c_j over rounds: $c_j = \sum_{r=t'}^{t} c_j(r) \leq \sum_{r=t'}^{t'+t-1} c_j(r)$. To simplify notation we relabel round t' round 1 and so round $t'+t-1$ becomes round t. Expanding c_j^k out fully using the summation:

$$\mathbb{E}\left[c_j^k\right] = \mathbb{E}\left[\sum_{r_1=1}^{t}\sum_{r_2=1}^{t}...\sum_{r_k=1}^{t} c_j(r_1)c_j(r_2)...c_j(r_k)\right]$$
$$= \sum_{r_1=1}^{t}\sum_{r_2=1}^{t}...\sum_{r_k=1}^{t} \mathbb{E}\left[c_j(r_1)c_j(r_2)...c_j(r_k)\right].$$

$\mathbb{E}\left[c_j(r_1)c_j(r_2)...c_j(r_k)\right]$ is just the probability that the two agents collide in each of rounds $r_1, r_2, ...r_k$. Assume w.l.o.g. that $r_1 \le r_2 \le ... \le r_k$. By Lemma 3 this is:

$$O\left(\frac{1}{r_1(r_2-r_1+1)(r_3-r_2+1)...(r_k-r_{k-1}+1)}\right).$$

So we can rewrite, by linearity of expectation:

$$\mathbb{E}\left[c_j^k\right] = k! \sum_{r_1=1}^{t} ... \sum_{r_k=r_{k-1}}^{t} O\left(\frac{1}{r_1...(r_k-r_{k-1}+1)}\right).$$

We multiply by $k!$ since in this sum we only have *ordered* k-tuples, whereas the original sum is over *unordered* k-tuples. We can bound:

$$\sum_{r_k=r_{k-1}}^{t} \frac{1}{r_k - r_{k-1}+1} = 1 + \frac{1}{2} + ... + \frac{1}{t} = O(\log t)$$

so rearranging the sum and simplifying gives:

$$\mathbb{E}\left[c_j^k\right] = k! \sum_{r_1=1}^{t} \frac{1}{r_1} ... \sum_{r_k=r_{k-1}}^{t} \frac{1}{r_k - r_{k-1}+1}$$
$$= k! \sum_{r_1=1}^{t} \frac{1}{r_1} ... \sum_{r_{k-1}=r_{k-2}}^{t} \frac{1}{r_{k-2} - r_{k-1}+1} \cdot O(\log t).$$

We repeat this simplification for each level of summation replacing $\sum_{r_i=r_{i-1}}^{t} \frac{1}{r_i - r_{i-1}+1}$ with $O(\log t)$. Iterating through the k levels gives $\mathbb{E}\left[c_j^k\right] = O(k! \log^k t)$ giving the lemma. □

As with Lemma 3, the techniques used in Lemmas 5 and 6 can be applied to a single walk. We give two bounds that may be of independent interest (proofs in full paper)

Corollary 7 (Re-Visit Moment Bound). *Consider an agent a_1 randomly walking on a two-dimensional $\sqrt{A} \times \sqrt{A}$ torus that is initially located at a uniformly random location and takes $t \le A$ steps. Let c_j be the number of times a_1 visits node j. For $j \in [1, ...A]$ and all $k \ge 2$,*

$$\mathbb{E}\left[\bar{c}_j^k\right] = O\left(\frac{t}{A} \cdot k! \log^{k-1} t\right).$$

Corollary 8 (Equalization Moment Bound). *Consider an agent a_1 randomly walking on a two-dimensional $\sqrt{A} \times \sqrt{A}$ torus. If a_1 takes $t \le A$ steps and c is the number of times it returns to its starting position (the number of equalizations), for all $k \ge 2$, $\mathbb{E}\left[\bar{c}^k\right] = O\left(k! \log^k t\right)$.*

3.3 Concentration of Density Estimate

Armed with the moment bound of Lemma 6 we can finally 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 start by showing that $\sum_{j=1}^{n} \bar{c}_j$ is a sub-exponential random variable.

Corollary 9 ($\sum_{j=1}^{n} \bar{c}_j$ is sub-exponential). *Assuming $t \le A$, $\sum_{j=1}^{n} \bar{c}_j$ is sub-exponential with parameters $b = \Theta(\log t)$ and $\sigma^2 = \Theta(td \log t)$. Specifically, for any λ with $|\lambda| < \frac{1}{b}$ $\mathbb{E}\left[e^{\lambda \sum_{j=1}^{n} \bar{c}_j}\right] \le e^{\frac{\sigma^2 \lambda^2}{2}}$.*

Proof. By Lemma 6, for $\sigma^2 = \Theta(\frac{t \log t}{A})$ and $b = \Theta(\log t)$, $\bar{c}_j$ satisfies the Bernstein condition: $\mathbb{E}\left[\bar{c}_j^k\right] \le \frac{1}{2}k!\sigma^2 b^{k-2}$. This implies that $\bar{c}_j$ is sub-exponential with parameters $\sigma^2 = \Theta(\frac{t \log t}{A})$ and $b = \Theta(\log t)$ (see [23], Chapter 2). Since each $\bar{c}_j$ is independent, this gives us, for all λ with $|\lambda| \le \frac{1}{b}$:

$$\mathbb{E}\left[e^{\lambda \sum_{j=1}^{n} \bar{c}_j}\right] = \prod_{j=1}^{n} \mathbb{E}\, e^{\lambda \bar{c}_j} \le e^{n \cdot \frac{\lambda^2}{2} \cdot \Theta\left(\frac{t \log t}{A}\right)} = e^{\Theta(\lambda^2 td \log t)}.$$

This completes the proof by the definition of a sub-exponential random variable. □

We finally apply a standard sub-exponential tail bound [23] to prove our main result.

Lemma 10 (Sub-exponential tail). *Suppose that X is sub-exponential with parameters (σ^2, b). Then, for any $\Delta \le \frac{\sigma^2}{b}$, $\mathbb{P}[|X - \mathbb{E}\, X| \ge \Delta] \le 2e^{-\frac{\Delta^2}{2\sigma^2}}$.*

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, $\tilde{d} = \frac{1}{t}\sum_{j=1}^{n} c_j$, 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 9 and Lemma 10:

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

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

4. MORE GENERAL TOPOLOGIES

We now extend our results to a broader set of graph topologies, demonstrating the generality of the local mixing analysis discussed above. We illustrate divergence between local and global mixing properties, which can have significant effects on random walk-based algorithms.

4.1 From Repeat Collision Bounds to Estimation Accuracy

Our proofs are largely independent of graph structure, using just a re-collision probability bound (Lemma 3) and the regularity (uniform node degrees) of the grid, so agents remain uniformly distributed in each round. Hence, extending our results to other regular graphs primarily involves obtaining re-collision probability bounds for these graphs.

We consider agents on A node graphs that execute analogously to Algorithm 1, stepping to a random neighbor in each round. Again, we focus on the multi-agent case but similar bounds (resembling Corollaries 7, 8) hold for single walks. We start with a general lemma, giving density estimation accuracy in terms of re-collision probability. The proof (see full paper) closely follows our grid analysis.

Lemma 11 (General Accuracy Bound). *Consider a regular graph with A nodes such that, if two randomly walking agents a_1 and a_2 collide in round r, for any $0 \leq m \leq t$, the probability that they collide again in round $r+m$ is $\Theta(\beta(m))$ for some non-increasing function $\beta(m)$. Let $B(t) \stackrel{\text{def}}{=} \sum_{m=0}^{t} \beta(m)$. After running for $t \leq A$ steps, 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 $\epsilon = O\left(\sqrt{\frac{\log(1/\delta)B(t)}{td}}\right)$.*

Note that in the special case of the grid, by Lemma 3, we can set $\beta(m) = 1/(m+1)$ and hence $B(t) = \Theta(\log t)$, yielding Theorem 1.

Applying the above bound requires a constant factor approximation to the re-collision probability – the probability is $\Theta(\beta(m))$. Sometimes however, it is much easier to give just an upper bound – so the probability is $O(\beta(m))$. In this case a slightly weaker bound holds:

Lemma 12 (General Accuracy Bound 2). *Consider a regular graph with A nodes such that, if two randomly walking agents a_1 and a_2 collide in round r, for any $0 \leq m \leq t$, the probability that they collide again in round $r+m$ is $O(\beta(m))$ for some non-increasing function $\beta(m)$. Let $B(t) \stackrel{\text{def}}{=} \sum_{m=0}^{t} \beta(m)$. After running for $t \leq A$ steps, 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 $\epsilon = O\left(\sqrt{\frac{\log(1/\delta)\cdot B(t)^2}{td}}\right)$.*

4.2 k-Dimensional Tori

We consider general k-dimensional tori. As k increases, local mixing becomes stronger, fewer re-collisions occur, and density estimation becomes easier. For $k \geq 3$, although the torus still mixes slowly, density estimation is as accurate as on the complete graph! We first study the ring:

Lemma 13 (Re-collision Bound – Ring). *If two randomly walking agents a_1 and a_2 are located on a 1-dimensional torus (a ring) with A nodes, and collide in round r, for any $m \geq 0$, the probability that a_1 and a_2 collide again in round $r+m$ for $k \geq 1$ is $\Theta\left(\frac{1}{\sqrt{m+1}}\right) + O\left(\frac{1}{A}\right)$.*

Proof. We have already shown this re-collision bound in the proof of Lemma 3. It is identical to $\mathbb{P}[\mathcal{C}_x | M_x = m]$ on an $A \times A$ grid, which is bounded by $\Theta\left(\frac{1}{\sqrt{m+1}}\right) + O\left(\frac{1}{A}\right)$. □

For $m \leq A$, the $O\left(\frac{1}{A}\right)$ is absorbed into the $\Theta\left(\frac{1}{\sqrt{m+1}}\right)$ term. We estimate the sum of repeat collision probabilities using $\frac{1}{\sqrt{x}} \leq 2(\sqrt{x} - \sqrt{x-1})$ (derivation in full paper). So:

$$\sum_{m=0}^{t} \frac{1}{\sqrt{m+1}} \leq 1 + 2\left(\sqrt{2} - \sqrt{1}) + ... + (\sqrt{t+1} - \sqrt{t}\right)$$
$$= 2\sqrt{t+1} - 1.$$

Similarly, $\frac{1}{\sqrt{x-1}} \geq 2(\sqrt{x} - \sqrt{x-1})$ and so:

$$\sum_{m=0}^{t} \frac{1}{\sqrt{m+1}} \geq 2\left(\sqrt{2} - \sqrt{1}) + ... + (\sqrt{t+2} - \sqrt{t+1}\right)$$
$$= 2\sqrt{t+2} - 2.$$

So, overall $\sum_{m=0}^{t} \frac{1}{\sqrt{m+1}} = \Theta(\sqrt{t})$. Plugging into Lemma 11, on a ring, random walk-based density estimation gives: $\epsilon = O\left(\sqrt{\frac{\log(1/\delta)\sqrt{t}}{td}}\right) = O\left(\sqrt{\frac{\log(1/\delta)}{\sqrt{t}d}}\right)$. Rearranging, $t = \Theta\left(\left(\frac{\log(1/\delta)}{\epsilon^2 d}\right)^2\right)$ rounds are necessary to obtain a $1 \pm \epsilon$ approximation with probability $\geq 1-\delta$. Local mixing on the ring is much worse than on the torus– we expect to see $\Theta(\sqrt{t})$ rather than $\Theta(\log t)$ repeat collisions with every agent interacted with. Hence, density estimation is much more difficult, requiring t to be quadratic rather than linear in $1/d$.

4.2.1 Higher Dimensional Tori

We now cover $k \geq 3$. While global mixing time is on the order of $A^{2/k}$ [2] and so is slow if $k << A$, local mixing is so strong that our accuracy bounds actually match those of independent sampling! Throughout this section, we assume that k is a small constant and hide it in asymptotic notation.

Lemma 14 (Re-collision Bound – Torus). *If two randomly walking agents a_1 and a_2 are located on a k-dimensional torus with A nodes, and collide in round r, for any constant $k \geq 3$, $m \geq 0$, the probability that a_1 and a_2 collide in round $r+m$ is $\Theta\left(\frac{1}{(m+1)^{k/2}}\right) + O\left(\frac{1}{A}\right)$.*

Proof. We closely follow the proof of Lemma 3. a_1 and a_2 take $2m$ total steps: M_i in each dimension for $i \in [1, ..., k]$. Let $\mathcal{C}_i$ be the event that the agents have the same position in the i^{th} dimension in round $r+m$. Following Lemma 3,

$$\mathbb{P}[\mathcal{C}_i | M_i = m_i] = \Theta\left(\frac{1}{\sqrt{m_i+1}}\right) + O\left(\frac{1}{A^{1/k}}\right).$$

So,

$$\begin{aligned}
&\mathbb{P}[\mathcal{C} | M_1 = m_1, ..., M_k = m_k] \\
&\quad = \left[\Theta\left(\frac{1}{\sqrt{m_1+1}}\right) + O\left(\frac{1}{A^{1/k}}\right)\right] \cdot \\
&\quad .... \cdot \left[\Theta\left(\frac{1}{\sqrt{m_k+1}}\right) + O\left(\frac{1}{A^{1/k}}\right)\right]. \qquad (7)
\end{aligned}$$

In expectation, $M_i = 2m/k$. So by a Chernoff bound,

$$\mathbb{P}[M_i \leq m/k] \leq 2e^{-(1/2)^2 \cdot 2m/3k} = O\left(\frac{1}{(m+1)^{k/2}}\right)$$

again assuming k is a small constant. Union bounding over all k dimensions, we have $M_i \geq m/k$ for all i except with probability $O\left(\frac{1}{(m+1)^{k/2}}\right)$ and hence by (7):

$$\begin{aligned}
\mathbb{P}[\mathcal{C}] &= O\left(\frac{1}{(m+1)^{k/2}}\right) + \left[\Theta\left(\frac{1}{\sqrt{\frac{m}{k}+1}}\right) + O\left(\frac{1}{A^{\frac{1}{k}}}\right)\right]^k \\
&= \Theta\left(\frac{1}{(m+1)^{k/2}}\right) + O\left(\frac{1}{A}\right)
\end{aligned}$$

giving the lemma (again, asymptotic notation hides multiplicative factors in k since it is constant). □

We can plug the above bound into Lemma 12. For $t \leq A$ and $k \geq 3$, $\sum_{m=0}^{t}\left(\frac{1}{(m+1)^{k/2}} + \frac{1}{A}\right) < 1 + \sum_{m=0}^{\infty} \frac{1}{(m+1)^{k/2}} = O(1)$. So we can set $B(t) = 1$ and have $\epsilon = O\left(\frac{\sqrt{\log(1/\delta)}}{td}\right)$. Rearranging, we require $t = \Theta\left(\frac{\log(1/\delta)}{\epsilon^2 d}\right)$. This matches the performance of independent sampling up to constants.

4.3 Regular Expanders

When a graph *does* mix well globally, it also mixes well locally. The number of repeat collisions is low and accurate density estimation is possible. The most obvious example is the complete graph, on which random-walk based density estimation is equivalent to density estimation via independent sampling. We generalize to any regular expander.

LEMMA 15 (RE-COLLISION BOUND – EXPANDER). *Let G be a k-regular expander with A nodes and adjacency matrix $\mathbf{M}$. Let $\mathbf{W} = \frac{1}{k} \cdot \mathbf{M}$ be its random walk matrix, with eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_A$. Let $\lambda = \max\{|\lambda_2|, |\lambda_A|\}$. If two randomly walking agents a_1 and a_2 collide in round r, for any $m \geq 0$, the probability that they collide again in round $r + m$ is at most $\lambda^m + 2/A$.*

PROOF. Suppose that a_1 and a_2 collide at node i in round r. The probability they re-collide at round $r+m$ is $||\mathbf{W}^m \mathbf{e}_i||_2^2$, since for each j, $\mathbf{W}^m_{i,j} = (\mathbf{W}^m \mathbf{e}_i)_j$ is the probability an agent is at node j after round $r+m$ given that it is at node i after round r. We bound this norm using the following lemma on how rapidly an expander random walk converges to its stable distribution:

LEMMA 16 ([14]). *Let G be a k-regular expander with A nodes, adjacency matrix $\mathbf{M}$, and random walk matrix $\mathbf{W} = \frac{1}{k} \cdot \mathbf{M}$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_A$ be the eigenvalues of $\mathbf{W}$ and $\lambda = \max\{|\lambda_2|, |\lambda_A|\}$. For each $1 \leq j \leq n$,*

$$\left|(\mathbf{W}^m \cdot e_i)_j - \frac{1}{A}\right| \leq \lambda^m.$$

Now we can bound $||\mathbf{W}^m \mathbf{e}_i||_2^2$ by:

$$||\mathbf{W}^m \mathbf{e}_i||_2^2 = \sum_{j=1}^{A} (\mathbf{W}^m \mathbf{e}_i)_j^2 = \sum_{j=1}^{A} \left(\frac{1}{A} + \chi_j\right)^2$$

where $\chi_j \stackrel{\text{def}}{=} (\mathbf{W}^m \cdot \mathbf{e}_i)_j - \frac{1}{A}$ so that $\chi_j \in [-1/A, \lambda^m]$ by Lemma 16. Thus $\sum_j \chi_j = \sum_j (\mathbf{W}^m \mathbf{e}_i)_j - A \cdot (1/A) = 0$ and

$$\begin{aligned} ||\mathbf{W}^m \mathbf{e}_i||_2^2 &= \sum_{j=1}^{A} \left(\frac{1}{A} + \chi_j\right)^2 \\ &\leq \sum_{j=1}^{A} \left(\left(\frac{1}{A}\right)^2 + \frac{2\chi_j}{A} + \chi_j^2\right) = \frac{1}{A} + \sum_{j=1}^{A} \chi_j^2. \end{aligned}$$

$\sum_j \chi_j^2$ is maximized when the number of possible j with $\chi_j = \lambda^m$ is maximized. Let $S \subset [1, A]$ be the indices j with $\chi_j = \lambda^m$. Since $\sum_j \chi_j = 0$, we have $\sum_{j \in S} \lambda^m + \sum_{j \notin S} \chi_j = 0$. Therefore, $|S| \cdot \lambda^m \leq -\sum_{j \notin A} \chi_j \leq 1 - |S|/A$ and so $|S| \leq \frac{1}{\lambda^m + 1/A}$. Therefore,

$$\begin{aligned} \sum_{j=1}^{A} \chi_j^2 &\leq \sum_{j \in S} \lambda^{2m} + \sum_{j \notin S} \chi_j^2 \\ &\leq \frac{\lambda^{2m}}{\lambda^m + 1/A} + \frac{A - |S|}{A^2} \leq \lambda^m + 1/A. \end{aligned}$$

Thus, $||\mathbf{W}^m \mathbf{e}_i||_2^2 \leq \lambda^m + 2/A$, giving the lemma. □

We now apply Lemma 12, with $B(t) = \sum_{m=0}^{t} \beta(m) \leq \frac{1}{1-\lambda} + 2t/A$. Assuming $t = O(A)$,

$$\begin{aligned} \epsilon &= O\left(\sqrt{\frac{\log(1/\delta)(1/(1-\lambda) + 2t/A)^2}{td}}\right) \\ &= O\left(\sqrt{\frac{\log(1/\delta)}{td(1-\lambda)^2}}\right). \end{aligned}$$

Rearranging, $t = \Theta\left(\frac{\log(1/\delta)}{\epsilon^2 d(1-\lambda)^2}\right)$, matching independent sampling up to a factor of $O(1/(1-\lambda)^2)$.

4.4 k-Dimensional Hypercube

Finally, we give bounds for a k-dimensional hypercube. Such a graph has $A = 2^k$ vertices mapped to the elements of $\{\pm 1\}^k$, with an edge between any two vertices that differ by hamming distance 1. The hypercube is relatively fast mixing. Its adjacency matrix eigenvalues are $[-k, -k+2, -k+4, ..., k-4, k-2, k]$. Since it is bipartite, we can effectively ignore the negative eigenvalues and apply Lemma 15 with $\lambda = \Theta(1 - 2/k) = \Theta(1 - 1/\log A)$. This yields $t = \Theta\left(\frac{\log(1/\delta)\log^2(A)}{\epsilon^2 d}\right)$. However, it is possible to remove the dependence on A via a more refined analysis – while the global mixing time of the graph increases as A grows, local mixing becomes stronger!

LEMMA 17 (RE-COLLISION BOUND – HYPERCUBE). *If two randomly walking agents a_1 and a_2 are located on a k-dimensional hypercube with $A = 2^k$ vertices and collide in round r, for any $m \geq 0$, the probability that a_1 and a_2 collide in round $r + m$ is at most $(7/10)^m + O\left(\frac{1}{\sqrt{A}}\right)$.*

PROOF. A node of the hypercube can be represented as a k-bit string and each random walk step seen as choosing one of the bits uniformly at random and flipping it. If a_1 and a_2 collide, for each of the bit, the total number of times a_1 and a_2 chose that bit must be even. The total number of possible ways for re-collision to occur at round $r + m$ is exactly the number of ways $2m$ flips can be placed into k buckets, where each bucket has even number of elements:

$$\sum_{\substack{a_1 + \ldots + a_k = 2m \\ (a_i \mod 2) \equiv 0}} \frac{(2m)!}{a_1! \cdot \ldots \cdot a_k!}.$$

This value is equal to the coefficient of x^{2m} in the exponential generating function

$$\begin{aligned} &(2m)!\left(1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots\right)^k \\ &\qquad = (2m)!\left(\frac{e^x + e^{-x}}{2}\right)^k = \frac{(2m)!}{2^k} \sum_{i=0}^{k} \binom{k}{i} e^{x(2i-k)}. \end{aligned}$$

Differentiating $2m$ times, we see the coefficient of x^{2m} is:

$$\frac{1}{2^k} \sum_{i=0}^{k} \binom{k}{i} (2i-k)^{2m} = \sum_{i=0}^{k} \left(\binom{k}{i} / 2^k\right) \cdot (2i-k)^{2m}.$$

This summation is exactly $\mathbb{E}[X^{2m}]$, where X is a sum of k i.i.d. random variables each equal to 1 with probability 1/2 and -1 otherwise. For any $c \in (0,1]$, we can write:

$$\begin{aligned}\mathbb{E}[X^{2m}] &= \mathbb{E}[X^{2m} \mid |X| \geq ck] \cdot \mathbb{P}[|X| \geq ck] \\ &\quad + \mathbb{E}[X^{2m} \mid |X| \leq ck] \cdot \mathbb{P}[|X| \leq ck] \\ &\leq k^{2m}\, \mathbb{P}[|X| \geq ck] + (ck)^{2m}.\end{aligned}$$

To bound the return probability bound, we this count by the the total number of possible paths taken by a_1 and a_2 in m steps, k^{2m}, giving an upper bound of:

$$\mathbb{P}[|X| \geq ck] + c^{2m}.$$

By a Hoeffding bound, $\mathbb{P}[|X| \geq ck] \leq 2e^{-c^2k/2}$. If we set $c = \sqrt{\ln A/k} = \sqrt{\ln 2}$ then $\mathbb{P}[|X| \geq ck] \leq 1/\sqrt{A}$. So our final probability bound is:

$$\mathbb{P}[|X| \geq ck] + c^{2m} \leq \frac{1}{\sqrt{A}} + (\sqrt{\ln 2})^{2m} < \frac{1}{\sqrt{A}} + (7/10)^m.$$

Note that, by adjusting c, it is possible trade off the terms in the above bound, giving stronger inverse dependence on A at the expense of slower exponential decay in m. $\square$

Using Lemma 12, we have $B(t) = \sum_{m=0}^{t} \beta(m) \leq \frac{10}{3} + \frac{t}{\sqrt{A}}$. Assuming $t = O(\sqrt{A})$, this gives $\epsilon = O\left(\sqrt{\frac{\log(1/\delta)}{td}}\right)$ and so $t = \Theta\left(\frac{\log(1/\delta)}{\epsilon^2 d}\right)$, matching independent sampling.

5. APPLICATIONS

We conclude by discussing algorithmic applications of our ant-inspired density estimation algorithm (Algorithm 1), variations on this algorithm, and our analysis techniques.

5.1 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 using random walks [11, 12, 15, 16]. In these applications, 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 [17, 8]. One approach is to run a single random walk and count repeat node visits [15, 12]. Alternatively, [11] proposes running multiple random walks and counting their collisions. This can be significantly more efficient since the dominant cost is typically in link queries to the network. With multiple random walks, this cost can be trivially distributed to multiple servers.

Walks are first run for a 'burn-in' period so their locations are distributed approximately by the stable distribution of the network. 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 for network size.

We show that ant-inspired algorithms can give runtime improvements over this method. After burn-in, instead of halting the walks, we run them for *multiple rounds*, recording encounter rates as in Algorithm 1. This allows the use of fewer random walks, decreasing total burn-in cost, and leading to faster runtimes when mixing time is relatively slow, as is common in social network graphs [18].

5.1.1 Random Walk-Based Network Size Estimation

Consider an undirected, connected, non-bipartite graph $G = (V, E)$. Let S be the set of vertices of G that are 'known'. Initially, $S = \{v\}$ where v is a seed vertex. We can access G by looking up the neighborhood $\Gamma(v_i)$ of any vertex $v_i \in S$ and adding $\Gamma(v_i)$ to S.

To compute the number of nodes $|V|$ in the network, we could scan S, looking up the neighbors of each vertex and adding them to the set. After querying all nodes in S we will have $S = V$ and will know the network size. However, the number of queries required equals $|V|$. The goal is to estimate network size using a significantly more efficient random-walk based approach.

A number of challenges are introduced by this application. While we can simulate many random walks on G, we can no longer assume these random walks start at randomly chosen nodes, as we do not have the ability to uniformly sample nodes from the network. Instead, we must allow the random walks to run for a burn-in phase of length proportional to the mixing time of G. After this phase, the walks are distributed approximately according to the stable distribution of G.

In general, G is not regular. In the stable distribution, a random walk is located at a vertex with probability proportional to its degree. Hence, collisions tend to occur more at higher degree vertices. To correct for this bias, we count a collision at vertex v_i with weight $1/\deg(v_i)$. As we will see, with this modification, we must adjust our final estimate by the average degree of the graph, which we must estimate.

We use a natural generalization of re-collision probability. For any $i, j \in |V|$, let $p(v_i, v_j, m)$ be the probability of an m step random walk starting at v_i ending at v_j. Define:

$$\beta(m) \stackrel{\text{def}}{=} \frac{\max_{i,j} p(v_i, v_j, m)}{\deg(v_j)}$$

Intuitively, this is the maximum m step collision probability, weighted by degree since higher degree vertices are more likely to be visited in the stable distribution. Let $B(t) = \sum_{m=1}^{t} \beta(m)$. Note that this weighted $B(t)$ is upper bounded by the unweighted $B(t)$ used in Lemmas 11 and 12.

For simplicity, we ignore burn-in and assume that our walks start distributed exactly by the stable distribution of G. A walk starts at vertex v_i with probability $p_i \stackrel{\text{def}}{=} \frac{\deg(v_i)}{\sum_i \deg(v_i)} = \frac{\deg(v_i)}{2|E|}$ and initial locations are independent. We also assume knowledge of the average degree $\overline{\deg} = 2|E|/|V|$. In our full paper we rigorously analyze burn-in and show to estimate $\overline{\deg}$, completing our analysis.

Algorithm 2 Random Walk Network Size Estimation

input: step count t, average degree $\overline{\deg}$, n random starting locations $[w_1, ..., w_n]$ distributed according to the network's stable distribution

$[c_1, ..., c_n] := [0, 0, ..., 0]$
for $r = 1, ..., t$ **do**
  $\forall j$, set $w_j := randomElement(\Gamma(w_j))$
    ▷ $\Gamma(w_j)$ denotes the neighborhood of w_j.
  $\forall j$, set $c_j := c_j + \frac{count(w_j)}{\deg(w_j)}$
    ▷ $count(w_j)$ returns # other walkers at w_j.
end for
$C := \frac{\overline{\deg} \sum c_j}{\cdot n(n-1)t}$
return $\tilde{A} = 1/C$

Theorem 18. *If $n^2 t = \Theta\left(\frac{B(t)\overline{\deg}}{\epsilon^2 \delta}|V|\right)$, with probability $1-\delta$, Algorithm 2 returns $\tilde{A} \in [(1-\epsilon)|V|, (1+\epsilon)|V|]$.*

Throughout this section, we work directly with the weighted total collision count C, showing that it is close to its expectation with high probability and hence giving the accuracy bound for $\tilde{A}$. As in the density estimation case, we start by showing that C is correct in expectation.

LEMMA 19. $\mathbb{E}\, C = 1/|V|$.

PROOF. Let $c_j(r)$ be the number of collisions, weighted by inverse vertex degree, walk j expects to be involved in at round r. In each round all walks are at vertex v_i with probability $p_i = \frac{\deg(v_i)}{2|E|}$, so:

$$\mathbb{E}\, c_j(r) = \sum_{i=1}^{|V|} \left[\frac{\deg(v_i)}{2|E|} \cdot \frac{(n-1)\deg(v_i)}{2|E|} \cdot \frac{1}{\deg(v_i)} \right] = \frac{n-1}{4|E|^2} \sum_{i=1}^{|V|} \deg(v_i) = \frac{n-1}{2|E|}.$$

By linearity of expectation: $\mathbb{E}\, c_j = \frac{t(n-1)}{2|E|}$, $\mathbb{E} \sum c_j = \frac{tn(n-1)}{2|E|}$ and hence, $\mathbb{E}\, C = \frac{\overline{\deg}}{2|E|} = 1/|V|$. $\square$

We now need to show concentration of C about its expectation. Let $c_{i,j}$ be the weighted collision count between walks w_i and w_j where $i \neq j$. It is possible to closely follow the moment bound proof of Lemma 6 and show that $c_{i,j}$ is sub-exponential. However, unlike in the case of regular graphs, we will not be able to claim that the different $c_{i,j}$'s are independent. Hence, we will not be able to use the same sub-exponential tail bounds employed in Section 3.3.

Instead, we bound the second moment (the variance) of each $c_{i,j}$ and show concentration via Chebyshev's inequality. This leads to a linear rather than logarithmic dependence on the failure probability $1/\delta$. However, note that we can simply perform $\log(1/\delta)$ estimates each with failure probability $1/3$ and return the median, which will be correct with probability $1-\delta$. Variance proofs are deferred to our full paper, with the upshot being:

LEMMA 20 (TOTAL COLLISION VARIANCE BOUND). *Let* $\overline{C} = \frac{\overline{\deg} \sum_j \bar{c}_j}{n(n-1)t}$. $\mathbb{E}\left[\bar{C}^2\right] = O\left(\frac{1}{n^2 t} \cdot \frac{B(t)|E|}{|V|^2}\right)$.

With this variance bound, we can prove Theorem 18.

PROOF OF THEOREM 18. Note that $\bar{C} = C - \mathbb{E}\, C$. By Chebyshev's inequality Lemma 20 gives:

$$\mathbb{P}\left[|C - \mathbb{E}\, C| \geq \epsilon\, \mathbb{E}\, C\right] \leq \frac{1}{\epsilon^2 n^2 t} \cdot B(t)|E|.$$

Rearranging gives that, in order to have $\bar{C} \in \left[\frac{(1-\epsilon)}{|V|}, \frac{1+\epsilon}{|V|}\right]$ with probability δ, we must have:

$$n^2 t = \Theta\left(\frac{1}{\epsilon^2 \delta} B(t)|E|\right).$$

Assuming $\epsilon < 1/4$ this gives $\tilde{A} \in [(1-2\epsilon)|V|, (1+2\epsilon)|V|]$, giving the lemma after adjusting constants on ϵ. $\square$

5.1.2 Runtime and Comparision to Previous Work

Let $M = O\left(\frac{\log(|E|/\delta)}{1-\lambda}\right)$ denote the burn-in time required before running Algorithm 2 (see full paper for derivation). Ignoring average degree estimation, which is typically of lower order (see full paper), to obtain a $(1 \pm \epsilon)$ estimate of network size with probability $1-\delta$ we must run n random walks for $M + t$ steps, making $n(M+t)$ link queries, where by Theorem 18, $n = \Theta\left(\sqrt{\frac{|V| \cdot B(t) \overline{\deg}}{t \cdot \epsilon^2 \delta}}\right)$. In the special case with $t = 1$ we obtain a somewhat simpler bound (proof in full paper), requiring $n = \Theta\left(\sqrt{\frac{|V| \cdot \overline{\deg}}{\epsilon^2 \delta}}\right)$.

[11] also uses $t = 1$, but uses a different estimator tracking degrees, inverse degrees, and collisions. Roughly, they require: $n = \Theta\left(\max\left\{\sqrt{\frac{1}{\epsilon^2 \delta \sum p_i^2}}, \frac{\sum p_i^3}{\epsilon^2 \delta \left(\sum p_i^2\right)^2}\right\}\right)$ where $p_i = \frac{\deg(v_i)}{2|E|}$. Their first term can be rewritten as

$$\sqrt{\frac{2|E|}{\sum_{i=1}^{|V|} \deg(v_i)^2}} \cdot \sqrt{\frac{|V| \cdot \overline{\deg}}{\epsilon^2 \delta}}.$$

This will always be somewhat smaller than our bound term as $2|E| \leq \sum_{i=1}^{|V|} \deg(v_i)^2$. Their second term is harder to compare but is upper bounded by:

$$\frac{\sum p_i^3}{\epsilon^2 \delta \left(\sum p_i^2\right)^2} \leq \frac{\deg_{\max}^3}{\overline{\deg}^4 |V|^2} \cdot \frac{2|E|}{\epsilon^2 \delta} = \frac{\deg_{\max}^3}{\overline{\deg}^3} \cdot \frac{1}{\epsilon^2 \delta |V|}.$$

Assuming $\deg_{\max} / \overline{\deg}$ is not too large, this term will be small. However, a few very high degree nodes in an otherwise sparse graph can make it very large.

In sum, not directly comparable to [11], in the $t = 1$ case, assuming reasonable node degrees, our bounds are of the same order of magnitude. Further, the bound of Theorem 18 gives an important tradeoff for graphs with slow mixing time – we can increase the number of steps in our random walks, decreasing the total number of walks.

In our full paper we demonstrate that on a torus with ≥ 3 dimensions, our bounds give a polynomial speed up over [11]. We leave it as an open question to compare our bounds with those of [11] on more natural classes of graphs, and to determine either experimentally or theoretically, typical values of $B(t)$ on these graphs.

5.2 Robot Swarm Density Estimation

Algorithm 1 can be directly applied for simple and robust density estimation in robot swarms. Additionally, the algorithm can be used 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 prob $1 - 2\delta$, $\tilde{d}_P/\tilde{d} \in \left[\left(\frac{1-\epsilon}{1+\epsilon}\right) f_P, \left(\frac{1+\epsilon}{1-\epsilon}\right) f_P\right] = [(1 - O(\epsilon)) f_P, (1 + O(\epsilon)) f_P]$ for small ϵ.

In a biological setting, properties may include if an ant has recently completed a food foraging trip [9], or if an ant is a nestmate or enemy [1]. In a robotics setting, properties may include whether a robot is part of a certain task group, whether a robot has completed a task, or whether a robot has detected a certain event or environmental property.

5.3 Random Walk Sensor Network Sampling

Finally, we believe our moment bounds for a single random walk (Corollaries 7 and 8) can be applied to random walk-based distributed algorithms for sensor network sampling. We leave obtaining rigorous bounds to future work.

Random walk-based sensor network sampling [13, 3] is a technique in which a query message (a 'token') is initially sent by a base station to a 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 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 visits introduces computational burden – either the token message size must increase or nodes themselves must remember which tokens they have seen. We are hopeful that our moment 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.

6. ACKNOWLEDGEMENTS

We thank Amartya Shankha Biswas, Christopher Musco, and Mira Radeva for helpful discussions. This work was partially supported by NSF Grants BIO-1455983 and CCF-1461559, the NSF Center for Science of Information grant CCF-0939370, and AFOSR grant FA9550-13-1-0042.

7. REFERENCES

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

[2] D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs, 2002.

[3] C. Avin and C. 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.

[4] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher. Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified. *arXiv:1201.0559*, 2012.

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

[6] W. Feller. An introduction to probability theory and its applications. vol. i. 1950.

[7] D. Gillman. A Chernoff bound for random walks on expander graphs. *SIAM Journal on Computing*, 27(4):1203–1220, 1998.

[8] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. A walk in Facebook: Uniform sampling of users in online social networks. *arXiv:0906.0060*, 2009.

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

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

[11] L. Katzir, E. Liberty, and O. 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.

[12] M. Kurant, C. T. Butts, and A. Markopoulou. Graph size estimation. *arXiv:1210.0460*, 2012.

[13] L. Lima and J. 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.

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

[15] J. Lu and D. 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.

[16] J. Lu and H. Wang. Variance reduction in large graph sampling. *Information Processing & Management*, 50(3):476–491, 2014.

[17] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, pages 29–42. ACM, 2007.

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

[19] C. Musco, H.-H. Su, and N. Lynch. Ant-inspired density estimation via random walks. *arXiv:1603.02981*, 2016.

[20] S. C. Nicolis, G. Theraulaz, and J.-L. Deneubourg. The effect of aggregates on interaction rate in ant colonies. *Animal Behaviour*, 69(3):535–540, 2005.

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

[22] R. J. Schafer, S. Holmes, and D. M. Gordon. Forager activation and food availability in harvester ants. *Animal Behaviour*, 71(4):815–822, 2006.

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