LCA

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Population Size and Density Estimation

Farid Movahedi Naini (EPFL), Olivier Dousse (Nokia),
Patrick Thiran (EPFL), Martin Vetterli (EPFL).

CH-1015 Lausanne
Patrick.Thiran@epfl.ch
http://lca.epfl.ch

A Classical Problem

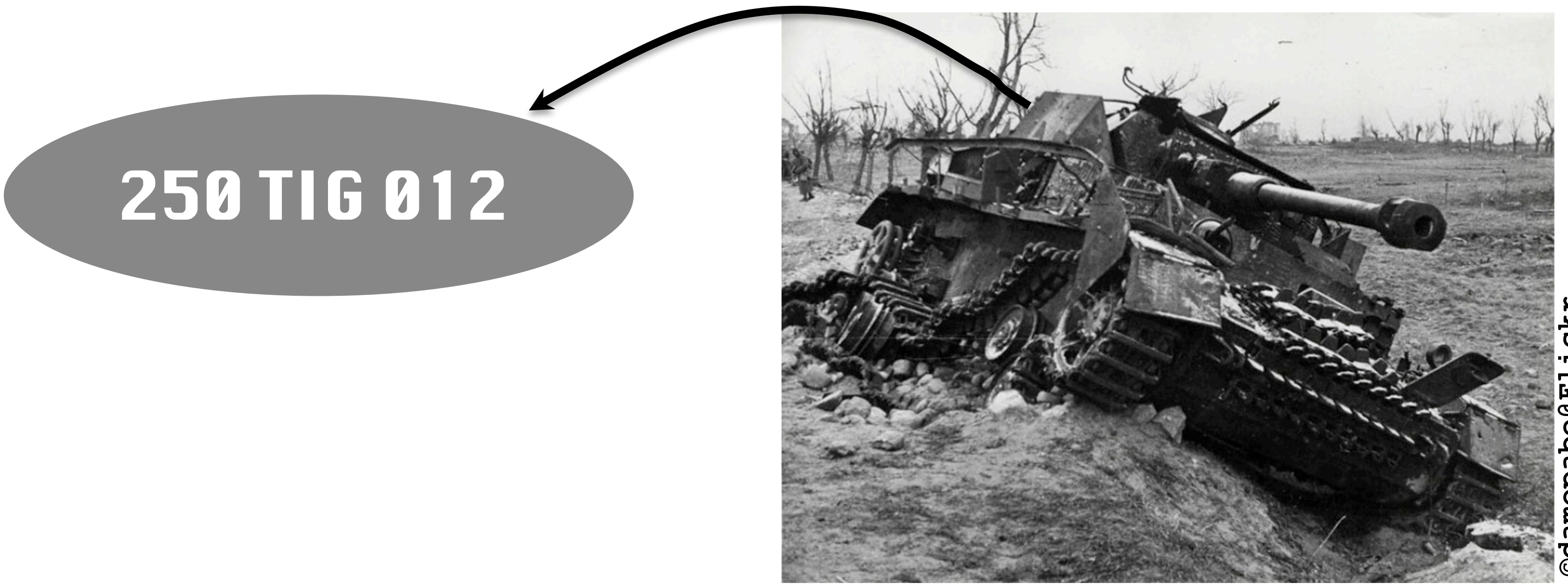
❑ Estimating Life's Diversity: How many species are there?

❑ Species/population estimation

- Biology: Estimating animal population size,
- Epidemiology: Estimating the number of drug users in a city,
- Information Theory: Alphabet size estimation.

Population Estimation: An Old Problem

- German Tank Problem: Population *N* of captured tanks.



- Minimum variance unbiased estimator:

$$\hat{N} = \max(serial\_nb) * (1 + 1/(sample\_size)) - 1$$

- August 1942 `(wikipedia)`
  - Intelligence: 1550
  - MVUE: 327
  - Groundtruth: 342.

Population Estimation Using Mobile Phones

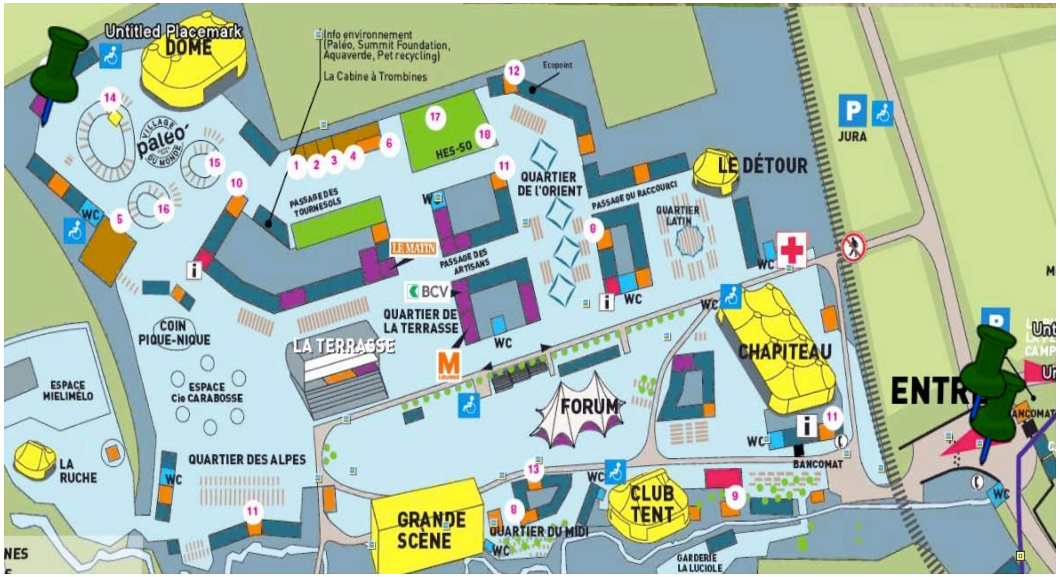
- Mobile network = distributed inference tool [NainiDTV14]
  - Mobile phones with Bluetooth and GPS.
- Broadcasts unique identifier in visible mode
  - Nominal range ~10 m.

Paléo Music Festival

- Major European music festival
  - July 20-25 2010, Nyon, Switzerland.
  - Attracts 40000 attendees per day.
  - An open-air environment (area 120000 m^2).

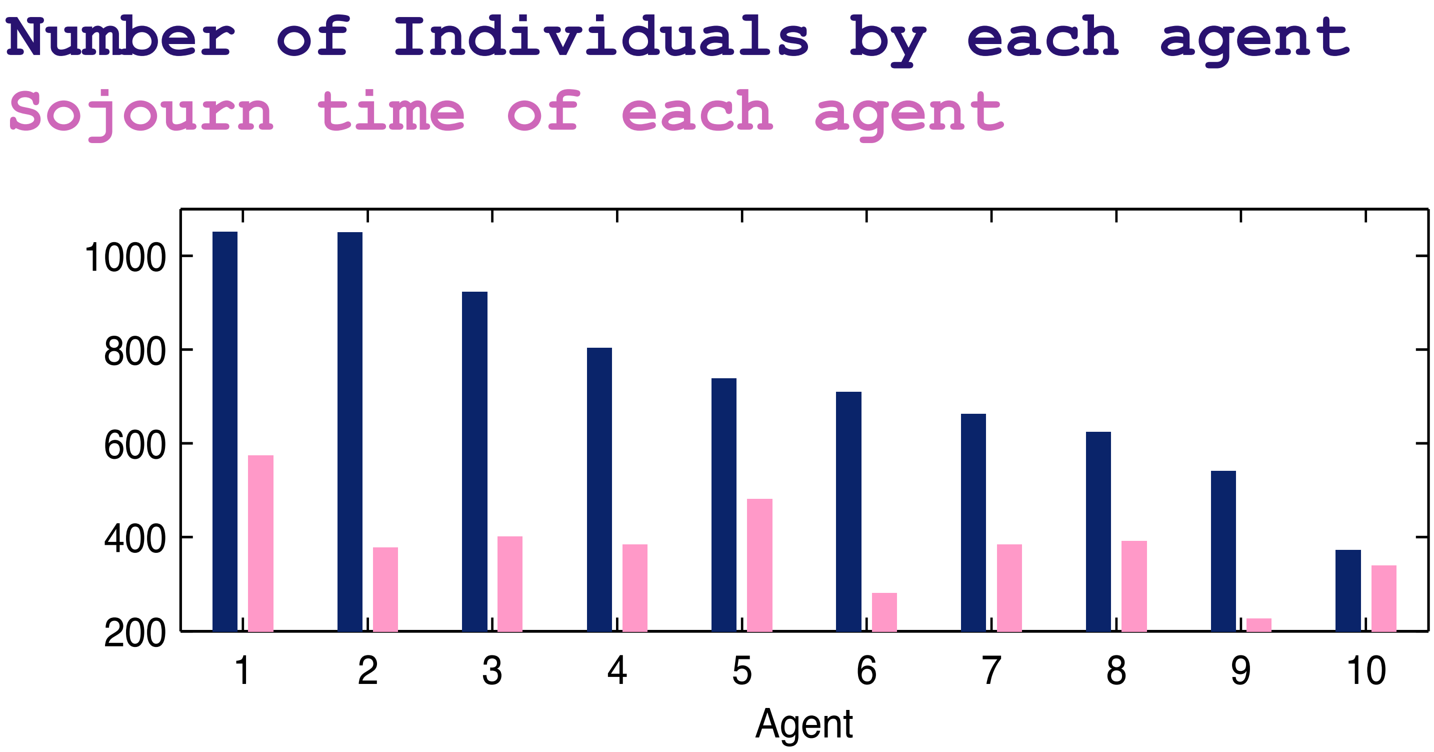
The Setup

- ❑ 10 arbitrary participants are sent to the festival:
  - Typical movement pattern of a participant.
  - Each carrying a Nokia N95 mobile phone.
- ❑ Three mobile phones installed at the entrances.
- ❑ All phones collect Bluetooth MAC addresses every 80 s.
- ❑ Data collected for one day of the festival (13 h).



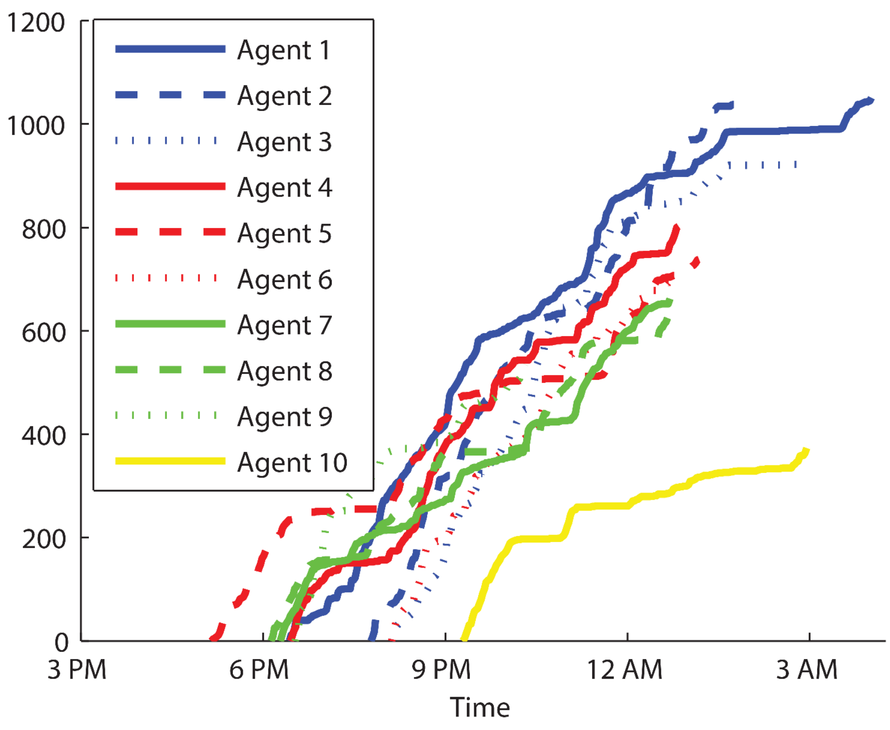
Coverage

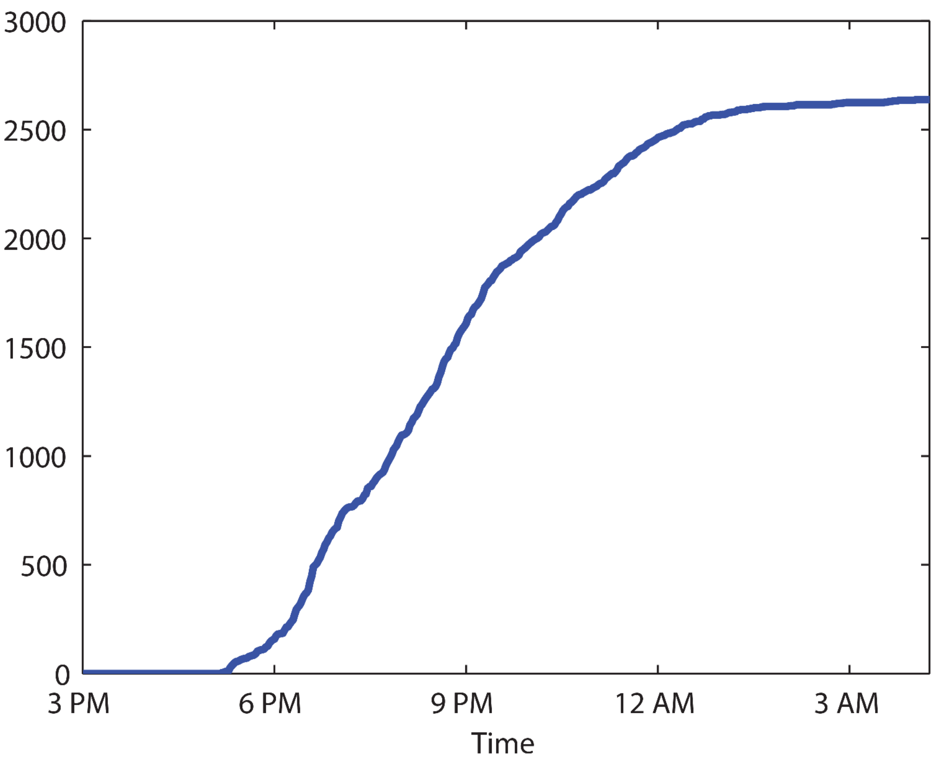
- 40536 attendees.
- M = 10 agents
- N = 3326 attendees with visible BT (8.2%).



Coverage

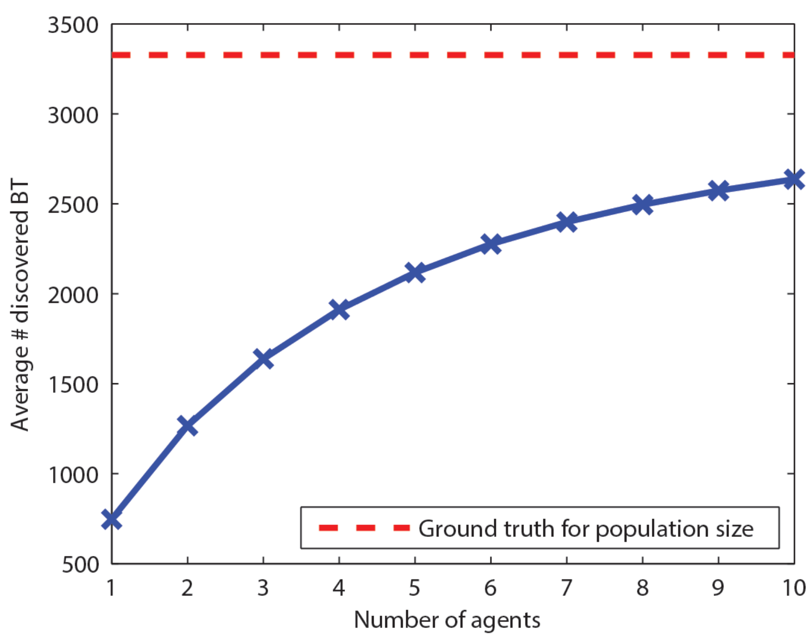
- 40536 attendees.
- N = 3326 attendees with visible BT (8.2%).
- Number of devices detected by mobile agents: n = 2637.
- 79.3% coverage (of visible BT) with only 10 agents.





Curve Fitting

- 2637 devices are detected ($N \geq 2637$): 20.7% undershoot.
- Actually, we have more fine-grain information:
  - Bluetooth traces of the $M = 10$ agents
  - Number k_{ij} of detections of individual i by agent j.
- Simple extrapolation = 2744 : 17.5% undershoot.
  - Averaged over subsets of m agents for $m = 1, 2, \ldots, 10$.



Use repetitions (capture-recapture)

- N distinct individuals,
- R_n = number of repeated individuals in sample of size n,
- $n_k = \min\{n : R_n = r\}$ (Here $n_1 = 4$, $n_2 = 5$, $n_3 = 7$),
- $N(n_k,k) \sim n_k^2 / (2r)$. [OrlitskySV, ISIT 2007]
- Assumes uniform i.i.d. sampling of the individuals.
- Here leads to $\hat{N} = 2676$
- Non uniform sampling of the individuals (N = 3326).

Pattern Maximum Likelihood (PML)

- ❑ Used for alphabet-size estimation [Acharya, Orlitsky, Pan et al]
- ❑ One source generating an i.i.d. sequence of symbols,
- ❑ Replace each symbol by its order of appearance → Pattern
- ❑ Example: 12311421
- ❑ Captures structure and frequencies, ignore symbols.
- ❑ Identify the distribution of the source that maximizes the probability of the observed pattern.

Pattern Maximum Likelihood (PML)

- Sequence maximum likelihood: which distribution maximizes the probability of the observed sequence?
  - Sequence of n distinct symbols.
  - Answer: Empirical frequency; alphabet size: n, each symbol probability $1/n$.
- Pattern maximum likelihood: which distribution maximizes the probability of the observed pattern?
  - Pattern: $123\ldots n$
  - Answer: large ($\gg n$).
  - Better model for estimating large alphabets from a small sample size.

Pattern maximum likelihood

- Obtaining the PML computationally expensive.
- Exact solution known for all patterns up to length *n* = 7.
- Expectation maximization (EM) algorithm for longer patterns, from [DhulipalaOS2003].
- For our experiment:
  - Input: number of contacts of each individual aggregated over all 10 agents (length: *n* = 11318).
  - Output: $\hat{N} = 3129$

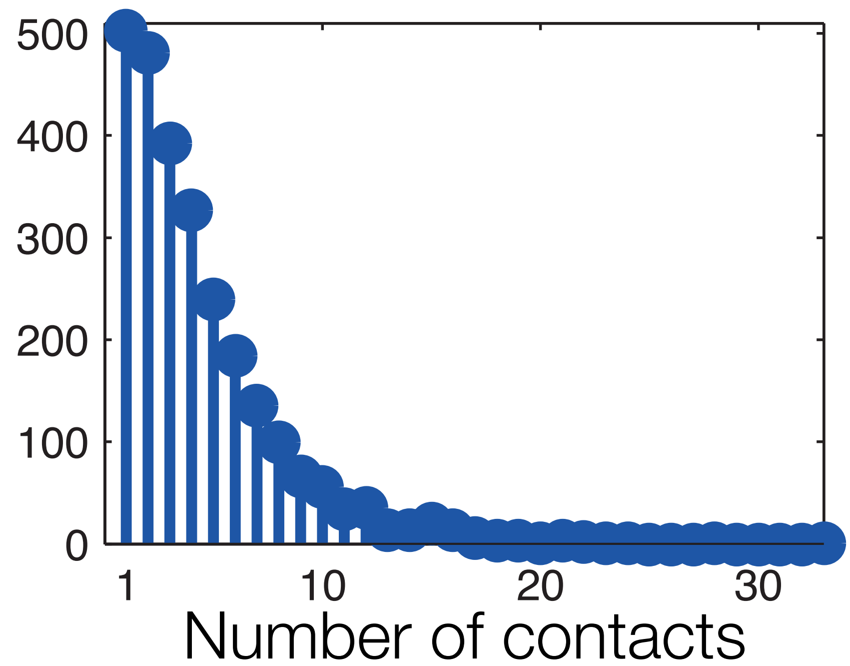
Opportunistic Mobile Sampling

- *M* agents > 1 source.
- Non-uniform random sampling.
- Time varying sampling.

Opportunistic Mobile Sampling

- M agents > 1 source.
- Non-uniform random sampling.
- Time varying sampling.

		Agent ID			
		1	2	j	$M = 10$
Individual ID	1	0	0		1
	2	0	2		0
	i			k_{ij}	
	2637	0	1		1



Parametric Model

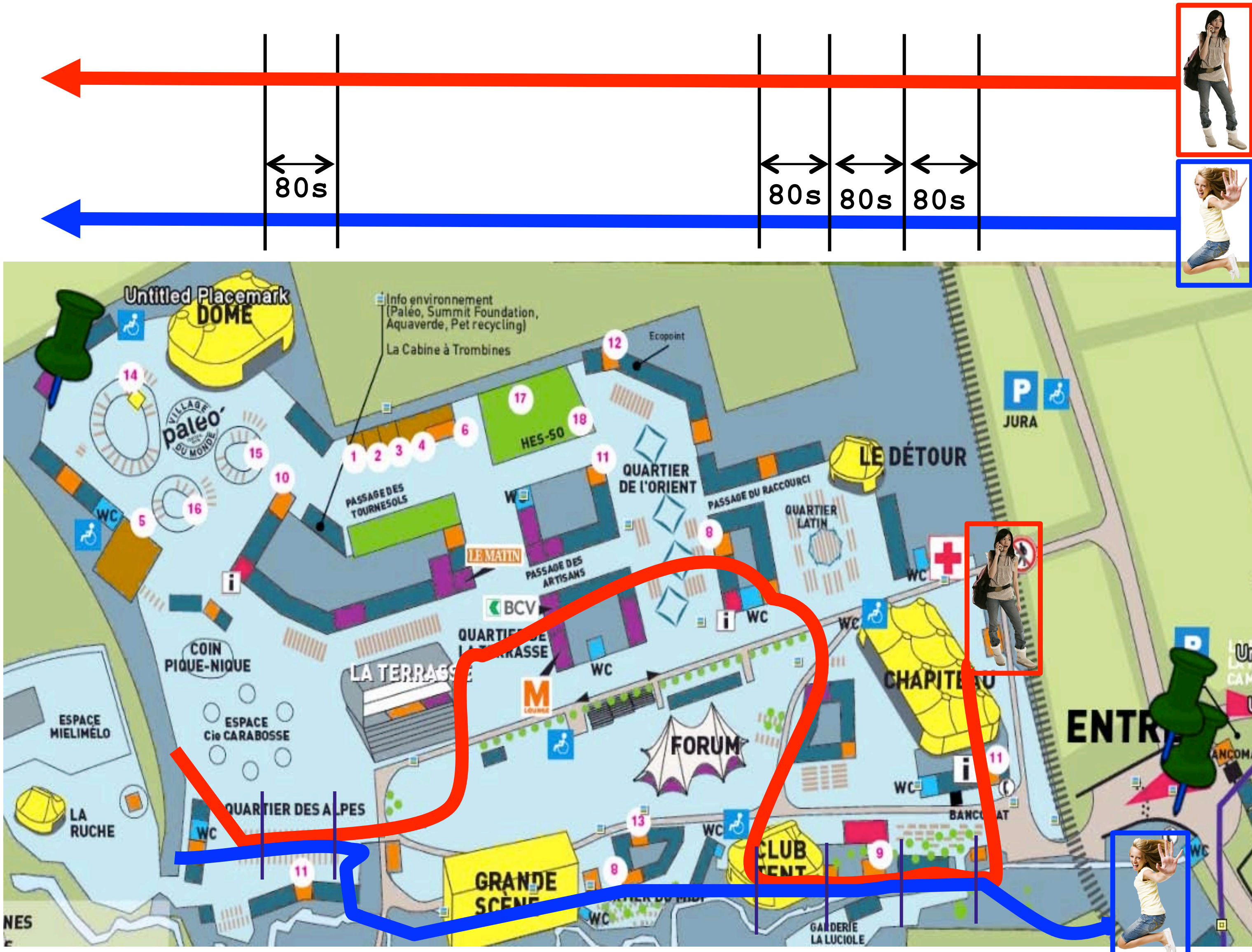
- Gamma-Poisson model:
  - Contacts are Poisson $k_{ij} \sim \text{Poisson}(\lambda_i)$

$$P(k_{ij} = k) = \frac{\lambda_i^k}{k!} \exp\left(-\lambda_i k\right)$$

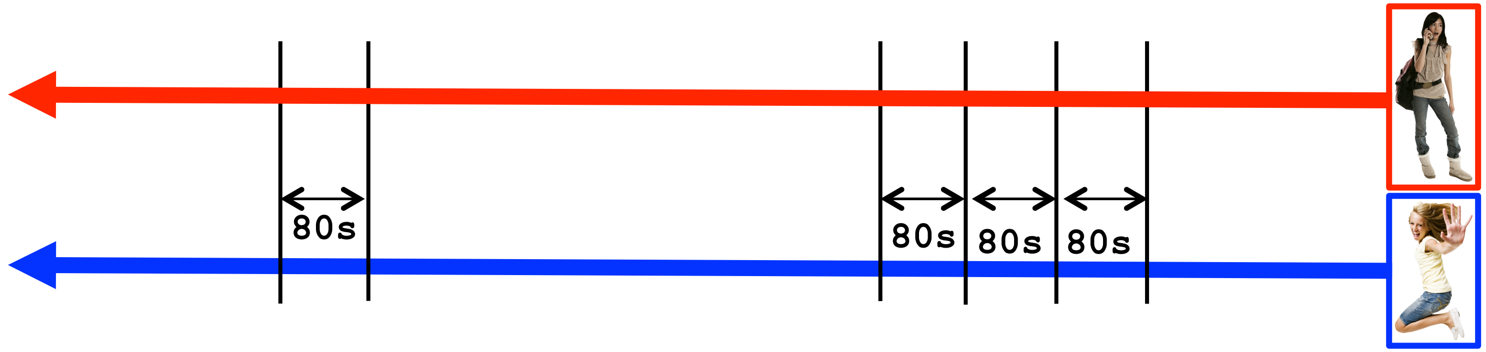
  - Gamma prior for the detection rate $\lambda_i \sim \Gamma(\alpha,\beta)$:

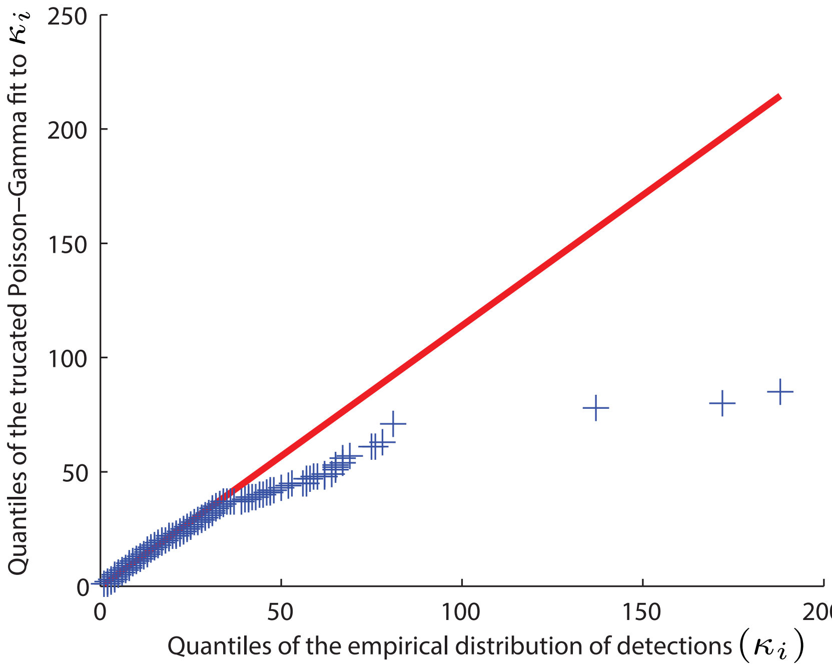
$$f_{\lambda_i}(\lambda) = \frac{\beta^\alpha}{\Gamma\left(\alpha\right)} \lambda^{\alpha-1} \exp\left(-\beta\lambda\right)$$

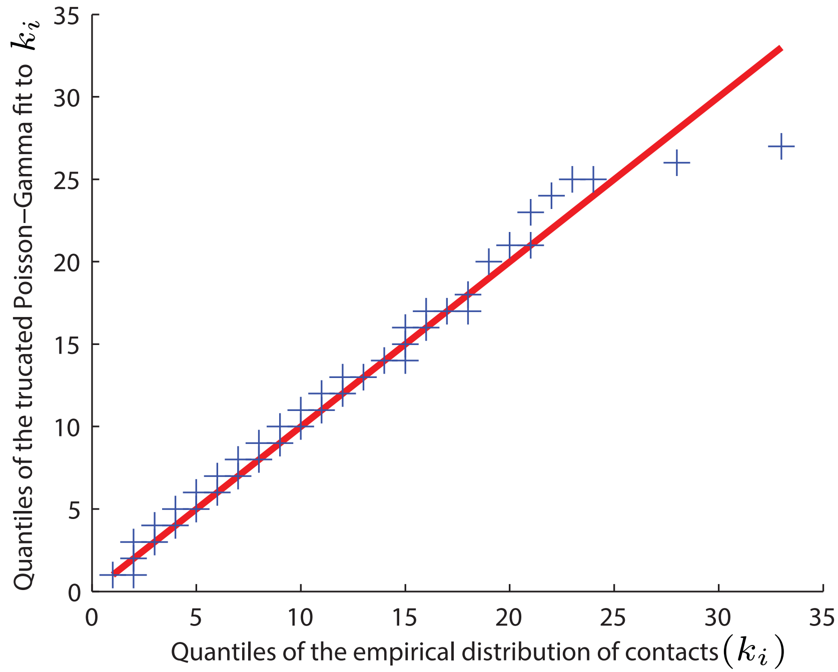
Detection vs Contact Times



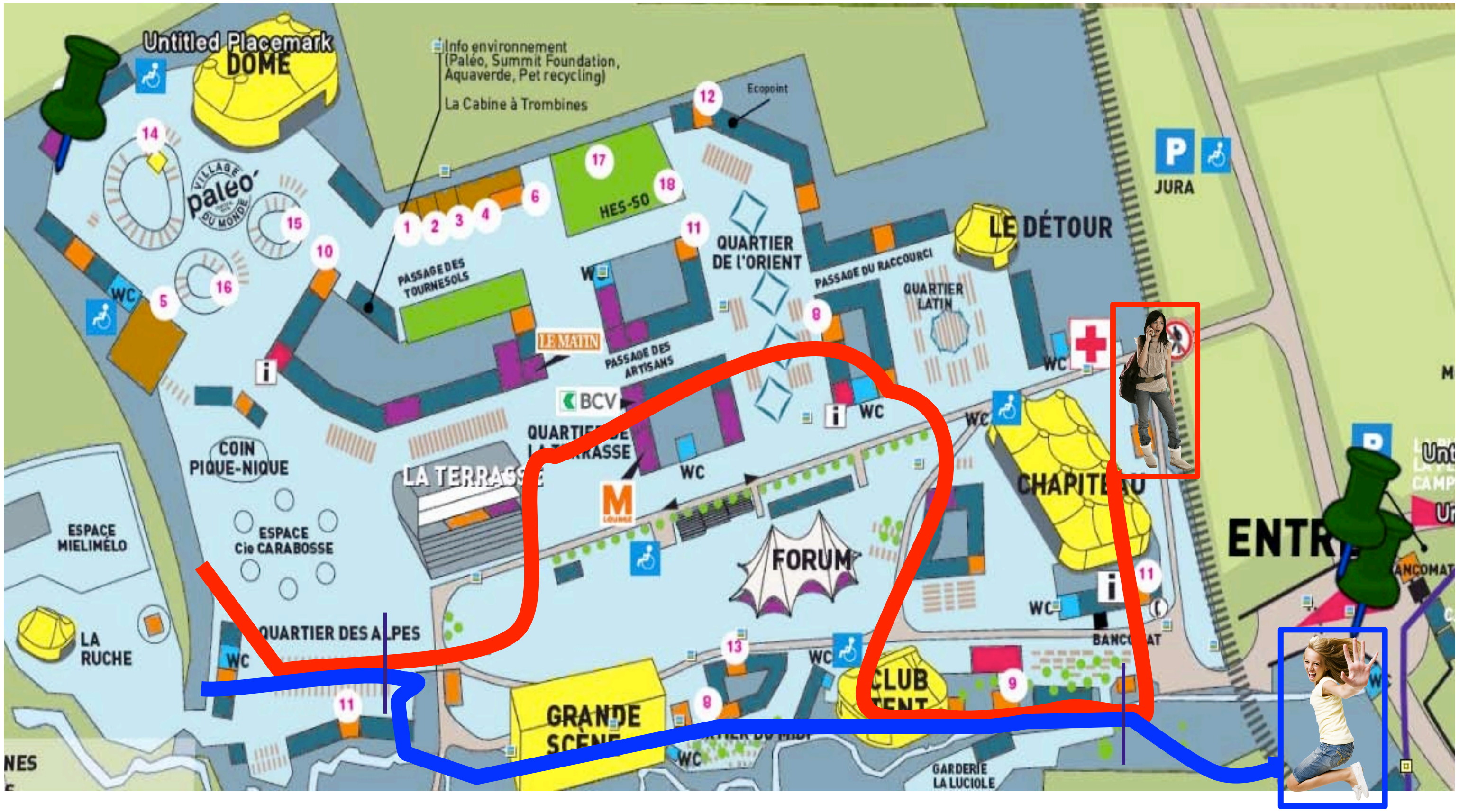
Detection vs Contact Times







Contact Times



Parametric Model

- Gamma-Poisson model:
  - Contacts are Poisson $k_{ij} \sim \text{Poisson}(\lambda_i)$

$$P(k_{ij} = k) = \frac{\lambda_i^k}{k!} \exp(-\lambda_i k)$$

  - Gamma prior for the detection rate $\lambda_i \sim \Gamma(\alpha,\beta)$:

$$f_{\lambda_i}(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\beta\lambda)$$

- Detection of individuals by agents are independent.

Parametric Model

- Device with detection rate $\lambda \sim \Gamma(\alpha,\beta)$
  - Probability that the individual be detected:

$$p_{det}^{(\lambda)} = 1 - \prod_{j=1}^{M} e^{-\lambda} = 1 - e^{-M\lambda}$$

$$p_{det}(\alpha,\beta) = \mathbb{E}_\lambda\left[p_{det}^{(\lambda)}\right] = 1 - \left(\frac{\beta}{\beta+M}\right)^{\alpha}$$

  - Probability that individual i is detected k_{i1} times by agent 1, …, k_{ij} times by agent j,…, k_{iM} times by agent M:

$$\mathsf{P}_i^{(\lambda)} = \prod_{j=1}^{M} e^{-\lambda}\frac{\lambda^{k_{ij}}}{k_{ij}!}$$

$$\mathsf{P}_i(\alpha,\beta) = \mathbb{E}_\lambda\left[\mathsf{P}_i^{(\lambda)}\right] = \frac{\Gamma(\alpha + \sum_{j=1}^{M} k_{ij})\beta^{\alpha}}{\Gamma(\alpha)(\beta+M)^{\alpha+\sum_{j=1}^{M} k_{ij}}}\prod_{j=1}^{M}\frac{1}{k_{ij}!},$$

Likelihood Based Estimator

- We maximize the likelihood function of the observation:

$$L(N,\alpha,\beta)=\boxed{\underbrace{\binom{N}{N-2637}\left(1-p_{det}(\alpha,\beta)\right)^{N-2637}}_{L_1(N,\alpha,\beta)}}\cdot\boxed{\underbrace{\prod_{i=1}^{2637}\mathsf{P}_i}_{L_2(\alpha,\beta)}}$$

- $L_1(N,\alpha,\beta)$: the likelihood of the unobserved individuals
- $L_2(\alpha,\beta)$: the likelihood of the observed individuals

$$L(N,\alpha,\beta)=\binom{N}{N-2633}\left(\frac{\beta}{\beta+M}\right)^{\alpha(N-2633)}\times\prod_{i=1}^{2633}\left\{\frac{\Gamma(\alpha+\sum_{j=1}^{M}k_{ij})\beta^{\alpha}}{\Gamma(\alpha)(\beta+M)^{\alpha+\sum_{j=1}^{M}k_{ij}}\prod_{j=1}^{M}k_{ij}!}\right\}$$

- We define the maximum likelihood estimators for *(N,α,β)*:

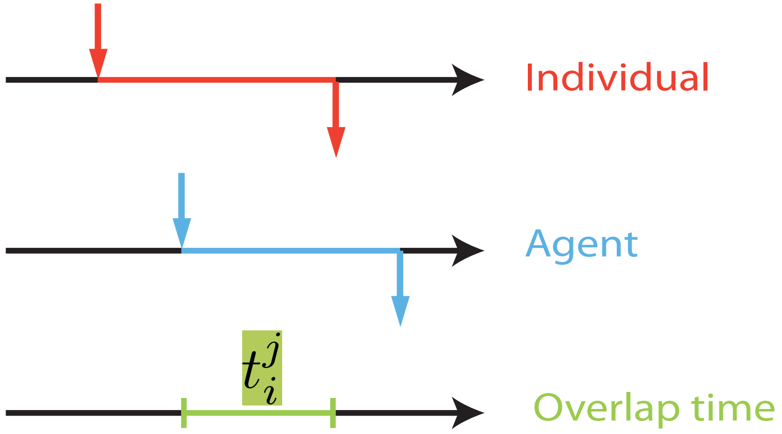
$$(\hat{N},\hat{\alpha},\hat{\beta})=\underset{N,\alpha,\beta}{\arg\max}\,\log L(N,\alpha,\beta)$$

Result

- Result of the MLE:

$\hat{N}$	$(N - \widehat{N})/N$
3106	6.61%

- Large undershoot
  - Attendees have different arrival/departure times
  - Assumed to be i.i.d.
- Overlap time between individual *i* and agent *j*'s



Contact intensity time-dependent

- Including arrival and departure times *at* and *dt* :
  - Overlap time $t_i^j = \min(dt_j, dt_i) - \max(at_j, at_i)$
  - (at_i, dt_i) known; (at_j, dt_j) estimated - joint distribution *f*.
  - $k_{ij} \sim Poisson(\lambda_i \cdot t_i^j)$
- The likelihood function has the same form:

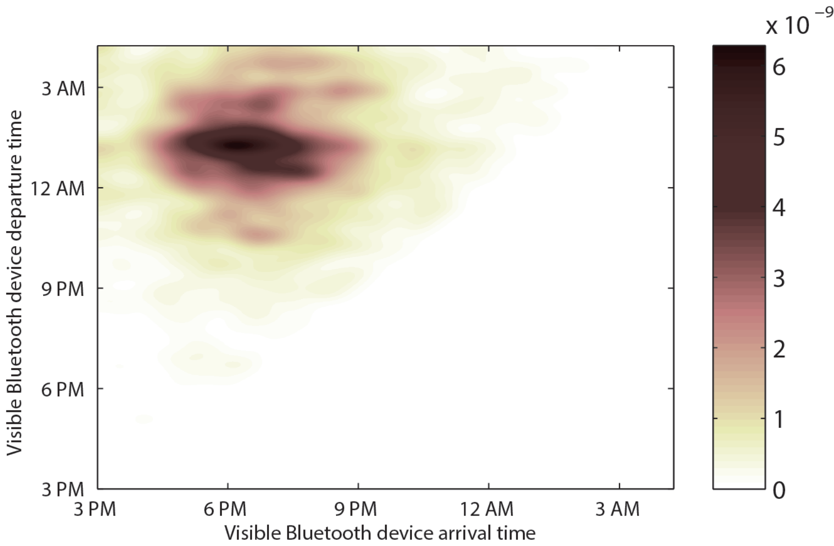
$$L(N,\alpha,\beta) = \underbrace{\binom{N}{N-2637}(1-p_{det}(\alpha,\beta))^{N-2637}}_{L_1(N,\alpha,\beta)} \cdot \underbrace{\prod_{i=1}^{2637} \mathsf{P}_i}_{L_2(\alpha,\beta)}$$

$$\mathsf{P}_i(\alpha,\beta) = \mathbb{E}_{f,\lambda}\left[\mathsf{P}_i^{(f,\lambda)}\right]$$

$$p_{det}(\alpha,\beta) = \mathbb{E}_{f,\lambda}\left[p_{det}^{(f,\lambda)}\right]$$

Result

- ❑ Distribution f(at,dt) of arrival/ departure times is measured or approximated by Gaussian



- ❑ Result of the MLE is:

Distribution of arrival/departures	$\hat{N}$	$(N - \widehat{N})/N$
Measured	3311	0.45%
Approximated	3275	1.53%

- ❑ Very small error
  - Gamma-Poisson model works well.
  - Inputs are minimally sufficient statistics for our MLE.

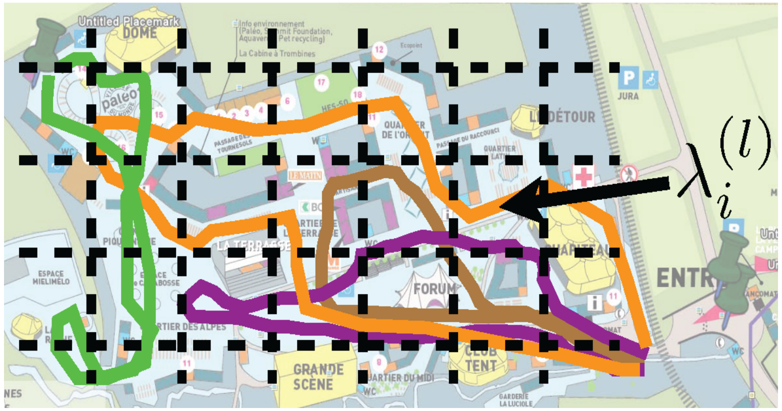
Results (N = 3326)

- We compare with two existing methods:

Method	$\hat{N}$	$(N - \widehat{N})/N$
Capture-recapture [LeeC1994]	3013	9.46%
Alphabet-size estimator [OrlitskySV2007]	2676	19.54%
PML [AcharyaOP09]	3129	5.95%
(*at*, *dt*) = maximal overlap (identical for all users *i*)	3106	6.61%
(*at*, *dt*) = measured	3311	0.45%
(*at*, *dt*) = Gaussian approximation	3275	1.53%

Population Density Estimation

- Divide area in *K* locations 1 ≤ *l* ≤ *K*
- Poisson contacts per location *l*:



$$k_{ij}^{(l)} \sim \mathrm{Poisson}(\lambda_i^{(l)} \cdot t_i^{j,(l)})$$

- $k_{ij}^{(l)}$ = number of times agent *j* contacts individual *i* in location *l*
- $t_i^{j,(l)}$ = overlap time between individual *i* and agent *j* in location *l*
- $\pi(l)$ = measures the density (popularity) of location l

$$\lambda_i^{(l)} \sim \Gamma(\pi(l)\alpha, \beta) \qquad \sum_{l=1}^{K} \pi(l) = 1$$

- Independence: $k_{ij}^{(l)} \perp\!\!\!\perp k_{i'j'}^{(l')}$ for *i* ≠ *i'*, *j* ≠ *j'* and/or *l* ≠ *l'*.

Likelihood Based Estimator

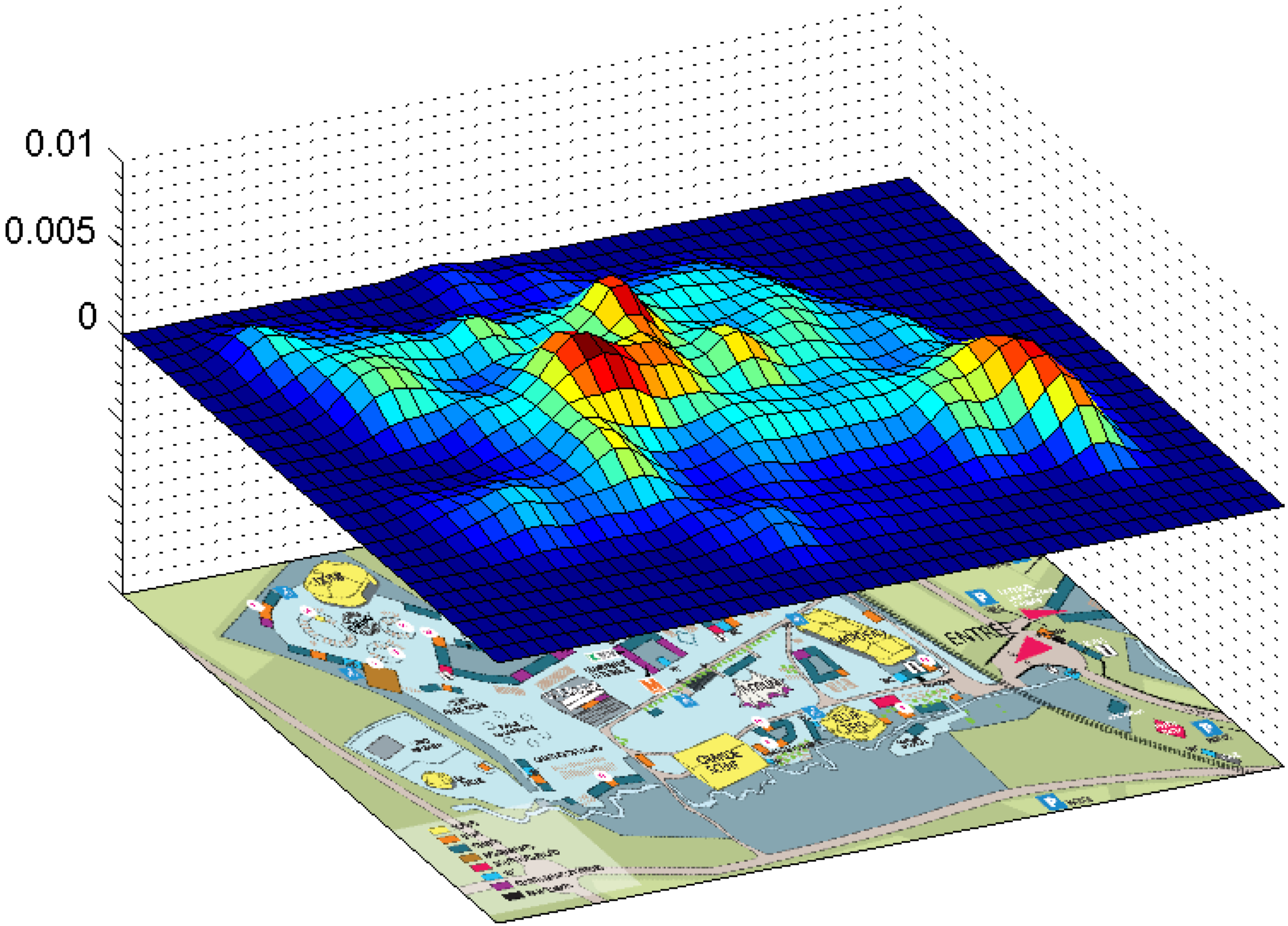
- Full likelihood function

$$L(N, \alpha, \beta, \pi(1), \pi(2), \ldots, \pi(K)) = \binom{N}{N-2637} (1 - p_{dsc}(\alpha, \beta))^{N-2637} \cdot \prod_{i=1}^{2637} \mathsf{P}_i$$

- Maximum likelihood estimator

$$\left(\widehat{N}, \widehat{\alpha}, \widehat{\beta}, \widehat{\pi}(1), \widehat{\pi}(2), \ldots, \widehat{\pi}(K)\right) = \underset{N,\alpha,\beta,\pi(1),\pi(2),\ldots,\pi(K)}{\arg\max} \log L(N, \alpha, \beta, \pi(1), \pi(2), \ldots, \pi(K))$$

Application to Paleo Festival

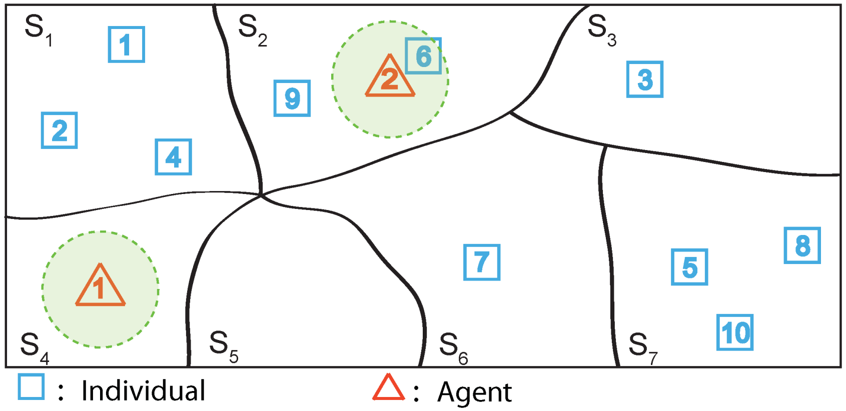


Impact of Mobility on Density Estimation

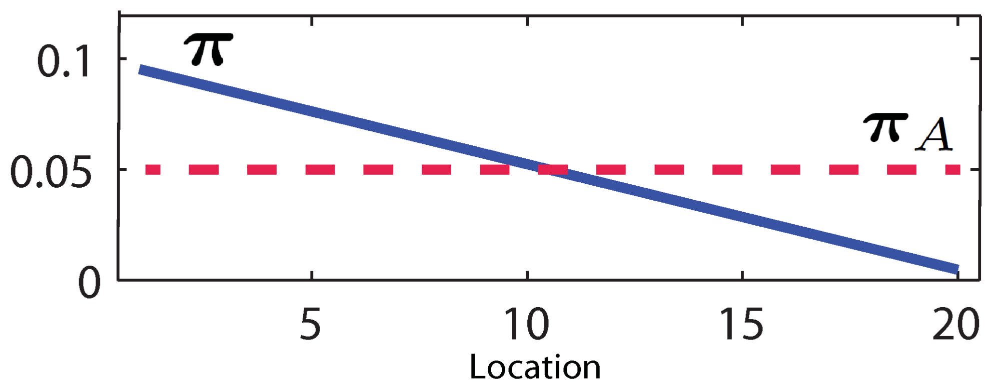
- How do mobile agents compare against static agents (e.g., sensors)?
- Methodology:
  - Simpler model for analytical tractability with explicit agents' mobility
    - Can quantitatively analyze the effect of agents' mobility
    - Can derive optimal random movement strategy for agents
  - Only estimation of density (Population size *N* known)
    - Can compute Fisher Information matrix for continuous parameters
    - Can analyze asymptotic behavior of parameter.

Discrete-time Model

- N known individuals, M agents moving between K locations

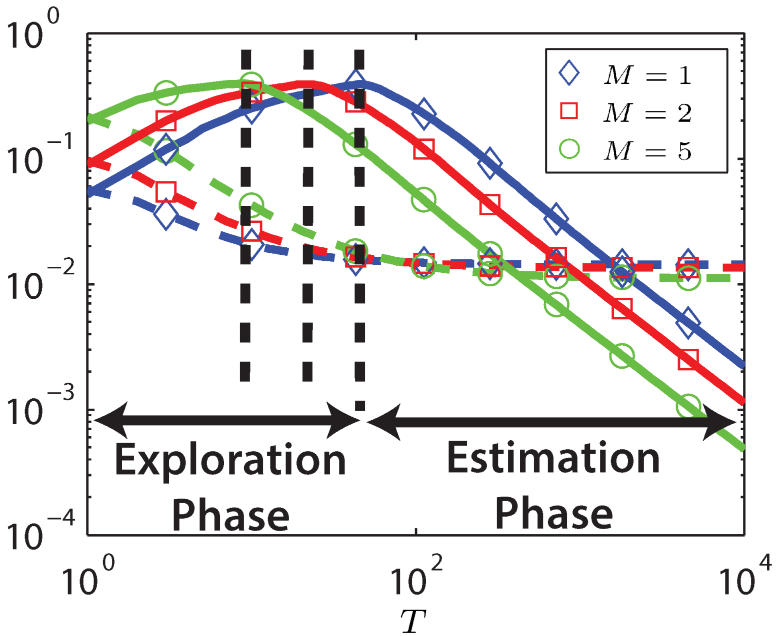


- At each time-sample $1 \le t \le T$, each individual and each agent choose a location i.i.d. according to π and π_A, respectively.
- Objective: Estimate π from agent's measurements.



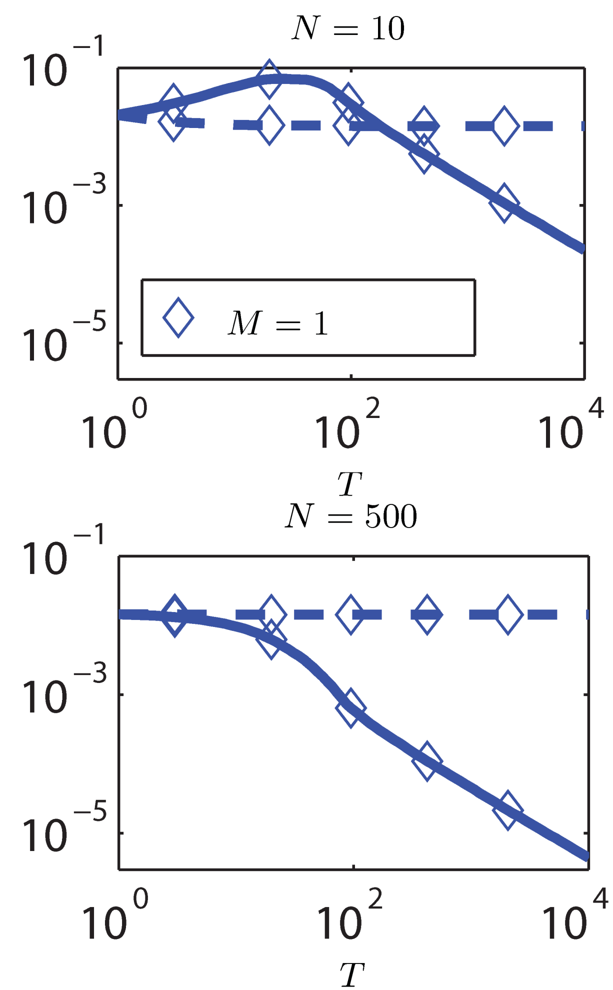
Simulation Results (K = 20 locations)

- Mobile vs static agents (N = 1)
  - Solid curve: mobile agents
  - Dashed curve: static agents



$$\mathrm{MSE} = \mathbb{E}\left[\|\boldsymbol{\pi} - \widehat{\boldsymbol{\pi}}_{MLE}\|_2^2\right]$$

- Mobile vs static agents (M = 1)



Conclusion

- Novel application that exploits the opportunistic contacts between mobile devices to infer population parameters
  - Focus on population size and density.
- The resulting estimate is surprisingly close to the ground truth
  - Considering the small number of agents,
  - But thanks to the large number of contacts.
- Exposure (overlap) time needs to be taken into account.
- Mobile agents outperform static agents for long observation intervals
  - Empirically verified for various sets of parameters.
  - Initial increase in the MSE theoretically shown for one particular scenario.

Acknowledgement

- The authors thank:
  - Stéphane Szilagyi, Pascal Viot (Paléo)
    - Help in cunducting the experiment at Paléo
  - Julien Eberle (EPFL)
    - Help in programming the mobile phones
  - Nokia Research Center
    - Research grant `Accidental Sampling’.