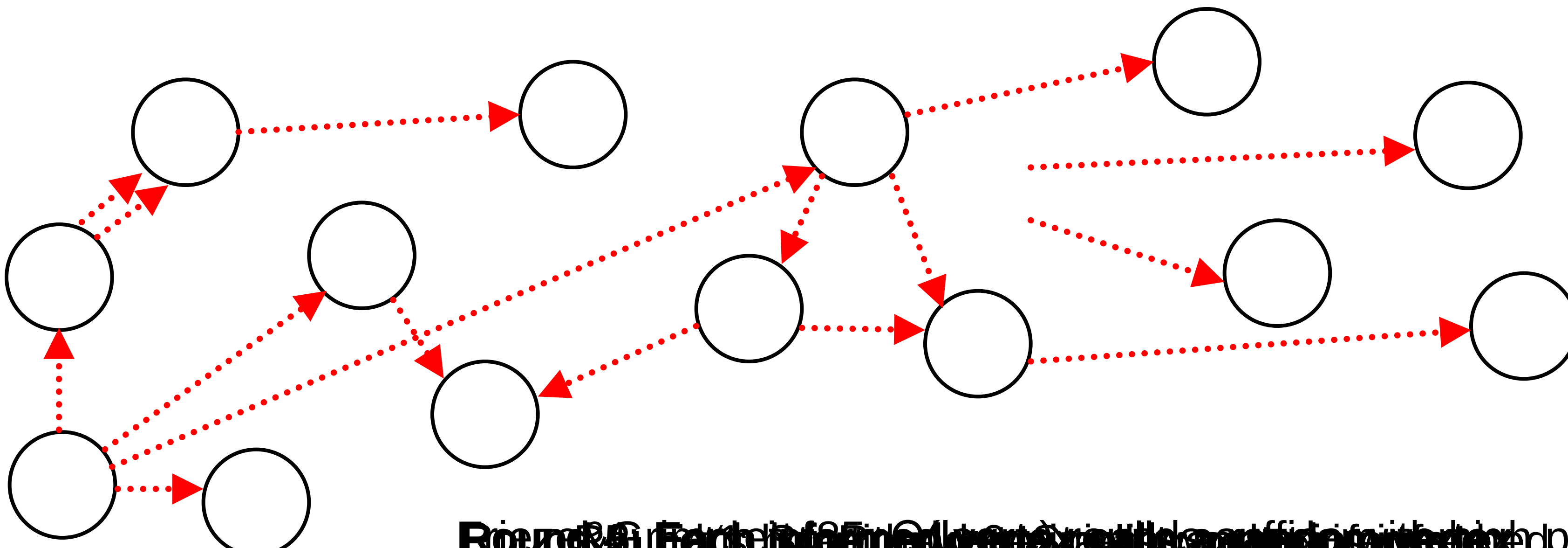
# Tight Analysis of Randomized Rumor Spreading in Complete Graphs

Benjamin Doerr (LIX)

joint work with Marvin Künnemann
(MPI Informatics)

# Randomized Rumor Spreading

- Randomized rumor spreading (here: in complete networks)
  - Start: One vertex knows a rumor
  - In each round, each vertex knowing the rumor calls a random vertex and forwards a copy of the rumor to it
  - Question: How long does it take until all vertices know the rumor?



# Why Study Rumor Spreading?

- Epidemic algorithm: Quickly and robustly spread information in a network!
  - Maintaining replicated databases: Name servers in the Xerox corporate internet [Dehmers et al. (1987)]
  - Communication in unreliable/unknown/dynamic/selfish networks
    - wireless sensor networks, mobile adhoc networks (robustness)
    - peer-to-peer sharing of large amounts of data (fairness, scalability)

- Model for existing processes
  - Rumors, computer viruses, diseases, influence processes, …

- An early motivation:
  - Technical tool in a mathematical analysis of an all-pairs shortest path algorithm [Frieze, Grimmett (1985)]

# Topic of This Talk: Rumor Spreading in Complete Graphs

- Most basic among all rumor spreading processes
- Only interest (in this talk): Runtime $S_n$ = time until all know the rumor

- Frieze&Grimmett'85:
  With probability $1 - o(1)$, $S_n = \log_2 n + \ln n \pm o(\log n)$.

- Pittel'87: Let $h = \omega(1)$ be arbitrary. Then
  with probability $1 - o(1)$, $S_n = \log_2 n + \ln n \pm o(h(n))$.

- Our result:
- $S_n$ is dominated by $\lceil \log_2 n \rceil + \frac{1}{n} CouponColl(n) + 2.6 + Geom(1 - o(1))$.
- $S_n$ is subdominated by $\lfloor \log_2 n \rfloor + \frac{1}{n} CouponColl\left(\frac{n}{2} \to n\right) - 1$.
- $\lfloor \log_2 n \rfloor + \ln n - 1.116 \leq E[S_n] \leq \lceil \log_2 n \rceil + \ln n + 2.765 + o(1)$.

# What is the Difficulty?

- Two different "regimes":

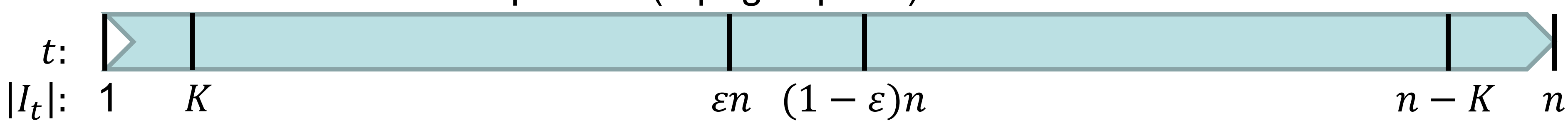
**Early: Informed nodes double.**

$$E[|I_{t+1}|] \geq 2\,|I_t|\left(1 - \frac{3|I_t|}{4n}\right)$$

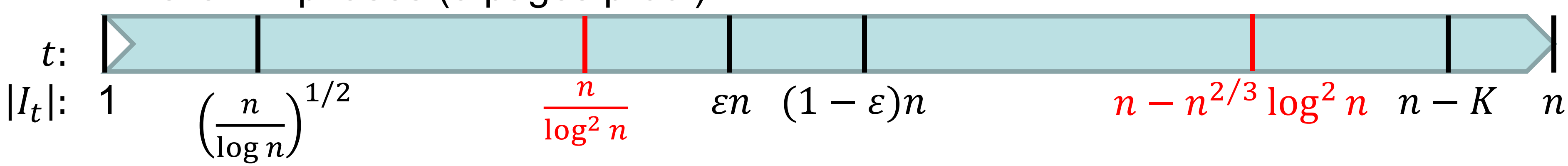
**Late: Uninformed shrink by $\approx 1/e$.**

$$E[|U_{t+1}|] \leq |U_t|\left(1 - \frac{1}{n}\right)^{n-|U_t|}$$

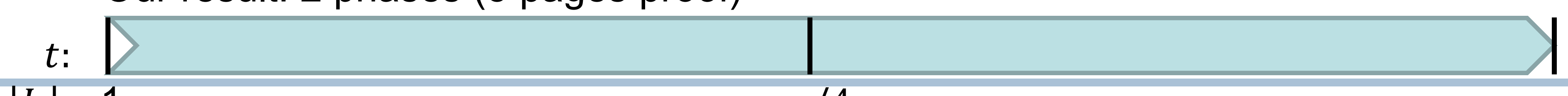
Frieze&Grimmett'85: 5 phases (7 pages proof)




Pittel'87: 7 phases (9 pages proof)




Our result: 2 phases (5 pages proof)



# The Early Regime

- Early regime: $E[|I_{t+1}|] \geq 2\,|I_t|\,\left(1 - \frac{3|I_t|}{4n}\right)$
- Chernoff bound: $|I_{t+1}| \geq 2\,|I_t|\,\left(1 - \frac{3|I_t|}{4n} - n^{-\varepsilon/2}\right)$ with prob. $1 - O(n^{-1+\varepsilon})$.
- Using $|I_t| \leq 2^t$ (ideal world): $|I_{t+1}| \geq 2\,|I_t|\,\left(1 - \frac{3}{4}\frac{2^t}{n} - n^{-\varepsilon/2}\right)$
- Then $|I_T| \geq 2^T \prod_{t=0}^{T-1}\left(1 - \frac{3}{4}\frac{2^t}{n} - n^{-\varepsilon/2}\right)$ [induction]
- $\geq 2^T \left(1 - \sum_{t=0}^{T-1}\left(\frac{3}{4}\frac{2^t}{n} + n^{-\varepsilon/2}\right)\right)$ [Bernoulli's inequality]
- $\geq 2^T \left(1 - \frac{3}{4}\frac{2^T}{n} - Tn^{-\varepsilon/2}\right)$, still with probability $1 - O(n^{-1+\varepsilon})$
- For $n = 2^k$: $|I_T| \geq \frac{5n}{16} - o(n)$ after $T = k - 1$ rounds.

Recall: $k-2$ rounds never give more than $|I_T| \leq n/4$.

# The Late Regime

- Late regime: $E[|U_{t+1}|] \le |U_t| \left(1 - \frac{1}{n}\right)^{n-|U_t|}$
  - If $|U_t| = cn$, then $E[|U_{t+1}|] \approx |U_t| e^{-1+c}$. (1)
  - Shrinking factor not constant → difficult ☹

- Solution: Detour via coupon collector!
  - (1) and induction and Chernoff: $|U_{t+i}| \le |U_t|(e^{-1+c+\varepsilon})^i$ very likely. [note: too weak to give a good run-time directly]
  - Total number of random calls in rounds $t+1 \dots t+l$:
    - $\sum_{i=1}^{l} |I_{t+i}| = \sum_{i=1}^{l}(n - |U_{t+i}|) \ge ln - |U_t| \sum_{i=1}^{l}(e^{-1+c+\varepsilon})^i = ln - O(n)$
    - [note: the $c$ and the $\varepsilon$ don't matter a lot]
  - If this number is larger than $CouponColl((1-c)n \to n)$, then all nodes are informed [so $l = \ln(n) + \Theta(1)$ does the job].
    - Our constant follow from $c = 5/16$ and $\varepsilon = o(1)$.

# Optimizing the Additive Constant

- Proof on last 2 slides reveals:
  - We do “perfect” work apart from the middle range where both $|I_t|$ and $|U_t|$ are $\Theta(n)$.
  - The biggest inaccuracies are caused by a handful of iterations before and after $|I_t| = \frac{5}{16}n$.
  - → Doing these by hand easily gives better bound ☺

- Karl Bringmann (MPI Informatics) did this numerically. Result:
  - $E[S_n] = \log_2 n + \ln n + 0.68252763\ldots$
  - what is hidden in the …, is not a constant, but a function of $n$ taking values in some interval of length less than $10^{-10}$.

Merci pour votre attention ☺