Web-graphs. Lecture 8.

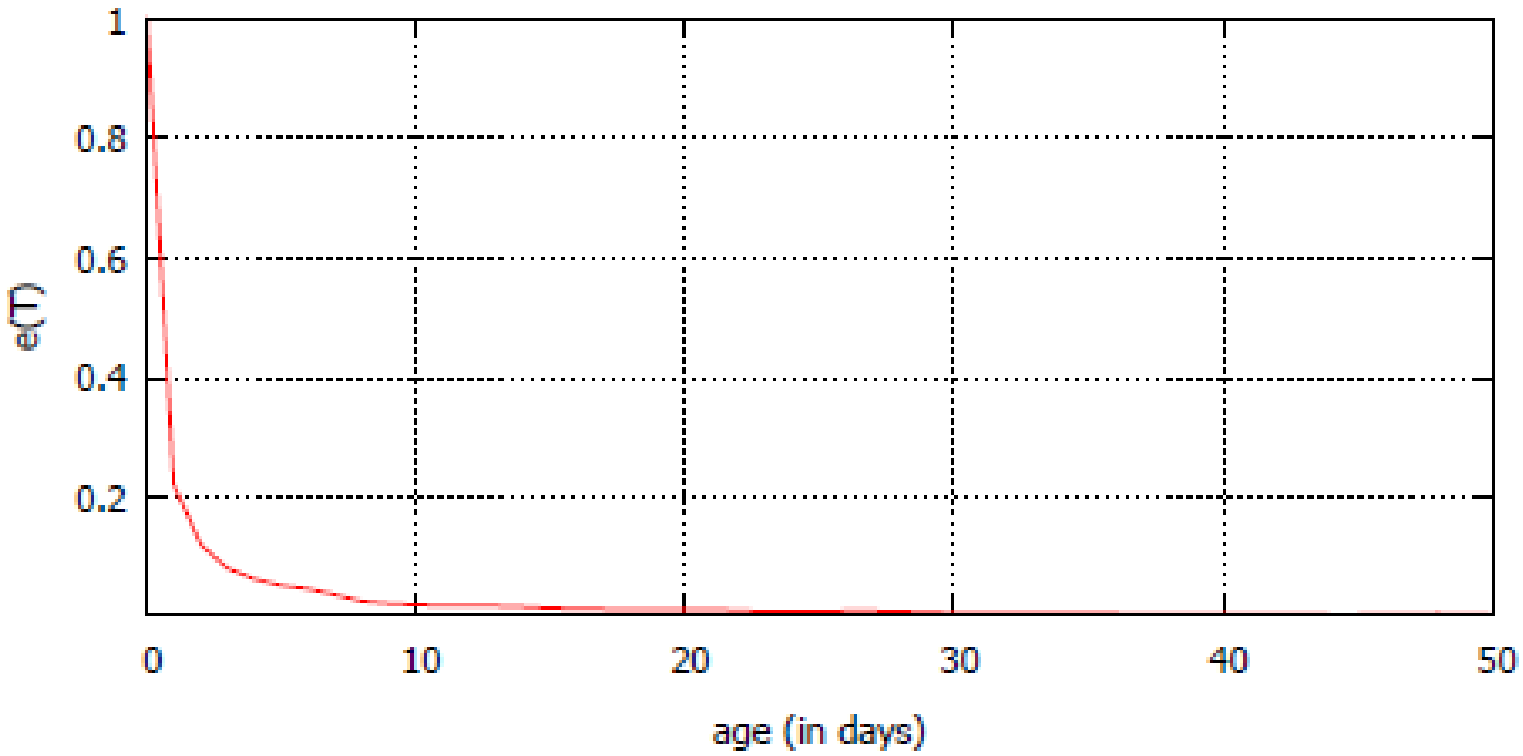
Freshness model

Based on: Lefortier, D., Ostroumova, L., & Samosvat, E. (2013, December). Evolution of the media web. In International Workshop on Algorithms and Models for the Web-Graph (pp. 80-92). Springer, Cham. (`https://arxiv.org/pdf/1209.4523.pdf`)

- One of the main drawbacks of the preferential attachment models, discussed previously, is that they pay **too much attention to old pages** and do not realistically explain how links pointing to newly-created pages appear
- Example: Media Web – the highly dynamic part of the Web related to media content where a lot of new pages appear daily.
- Most new media pages like news and blog posts are **popular only for a short period of time**, i.e., such pages are mostly cited and visited for several days after they appeared

Freshness model

- MemeTracker public data set, which covers 9 months of Media Web activity, was analyzed.
- 18M links and 6.5M documents (nodes)
- Let $e(T)$ be the fraction of edges connecting nodes whose age difference is greater than T.
- $e(T)$ for the dataset is decreasing exponentially fast, which is not the case for preferential attachment model



Freshness model

- Our aim is to build the graph sequence $\{G_n\}$
- m – the out-degree of each node
- $N(n)$ – integer function of n
- $\zeta_1, \zeta_2, \ldots$ – the sequence of collectively independent positive random values taken from the same distribution

Freshness model

- We firstly construct an auxiliary graph $\tilde{G}_t^n$
- $\tilde{G}_2^n$ consists of two nodes and one edge between them
- $q(1) = \zeta_1$ and $q(2) = \zeta_2$ are qualities of the first two nodes:
- At $t+1$ step $(2 \leq t \leq n-1)$ we add one new node and m links to the graph $\tilde{G}_t^n$
  - Quality of the new node $q(t+1) = \zeta_{t+1}$
  - $P(t+1 \to i) = \frac{\mathrm{attr}_t(i)}{\sum_{j=1}^{t} \mathrm{attr}_t(j)}$, function $\mathrm{attr}_t(i)$ is the node i's attractiveness at t time point
  - Function $\mathrm{attr}_t(i)$ is specified as follows:

$$\mathrm{attr}_t(i) = B_{it}^1 B_{it}^2 B_{it}^3$$

where

$$B_{it}^1 = (1 \text{ or } q(i))$$
$$B_{it}^2 = (1 \text{ or } d_t(i))$$
$$B_{it}^3 = (1 \text{ or } \mathbb{I}[i > t - N(n)] \text{ or } e^{-\frac{t-i}{N(n)}})$$

$d_t(i)$ is the node i degree in $\tilde{G}_t^n$

- $G_n = \tilde{G}_n^n$

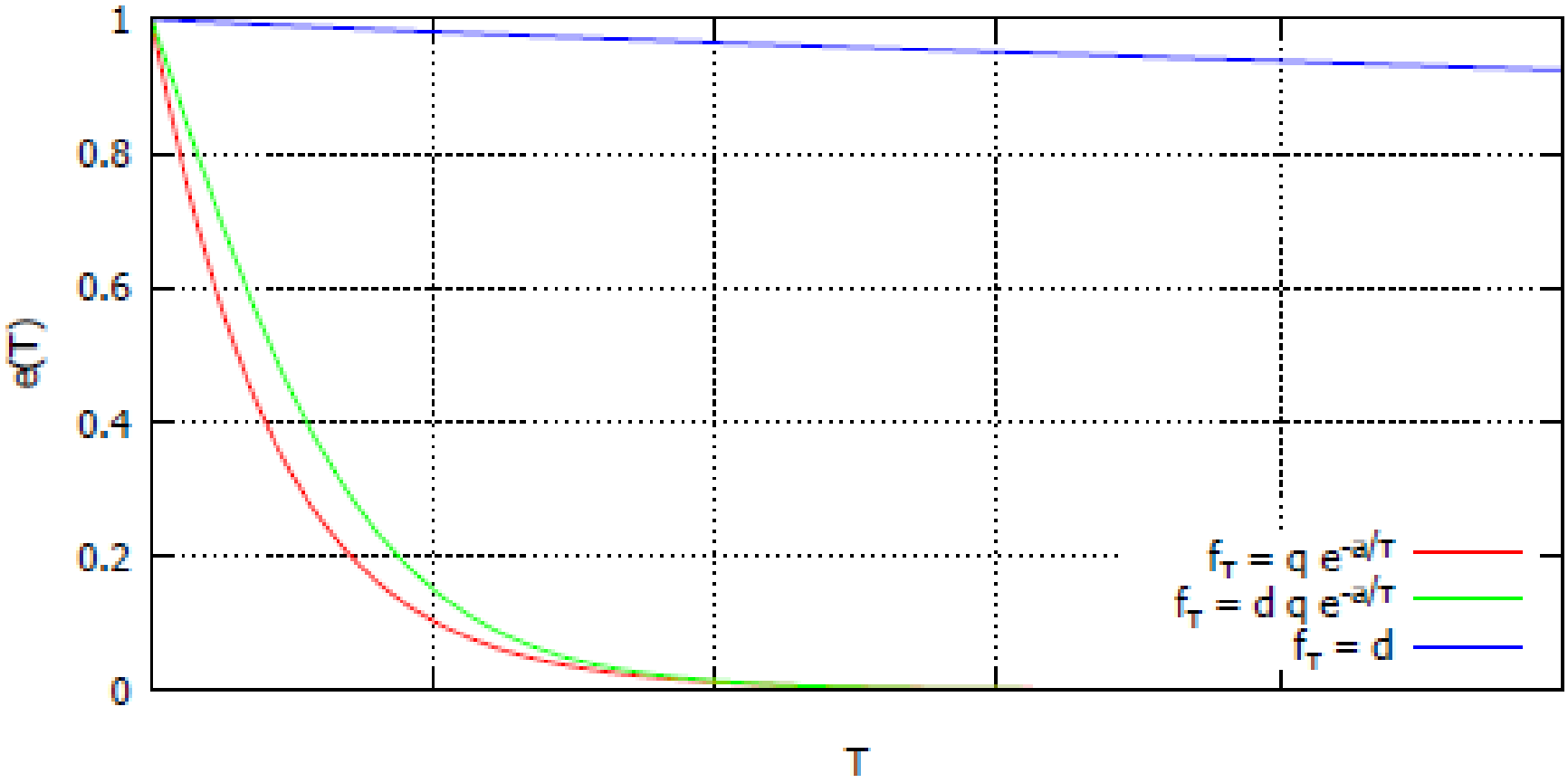
Freshness model

- The most realistic case is the case when

$$\mathrm{attr}_t(i) = q(i)e^{-\frac{t-i}{N(n)}}$$

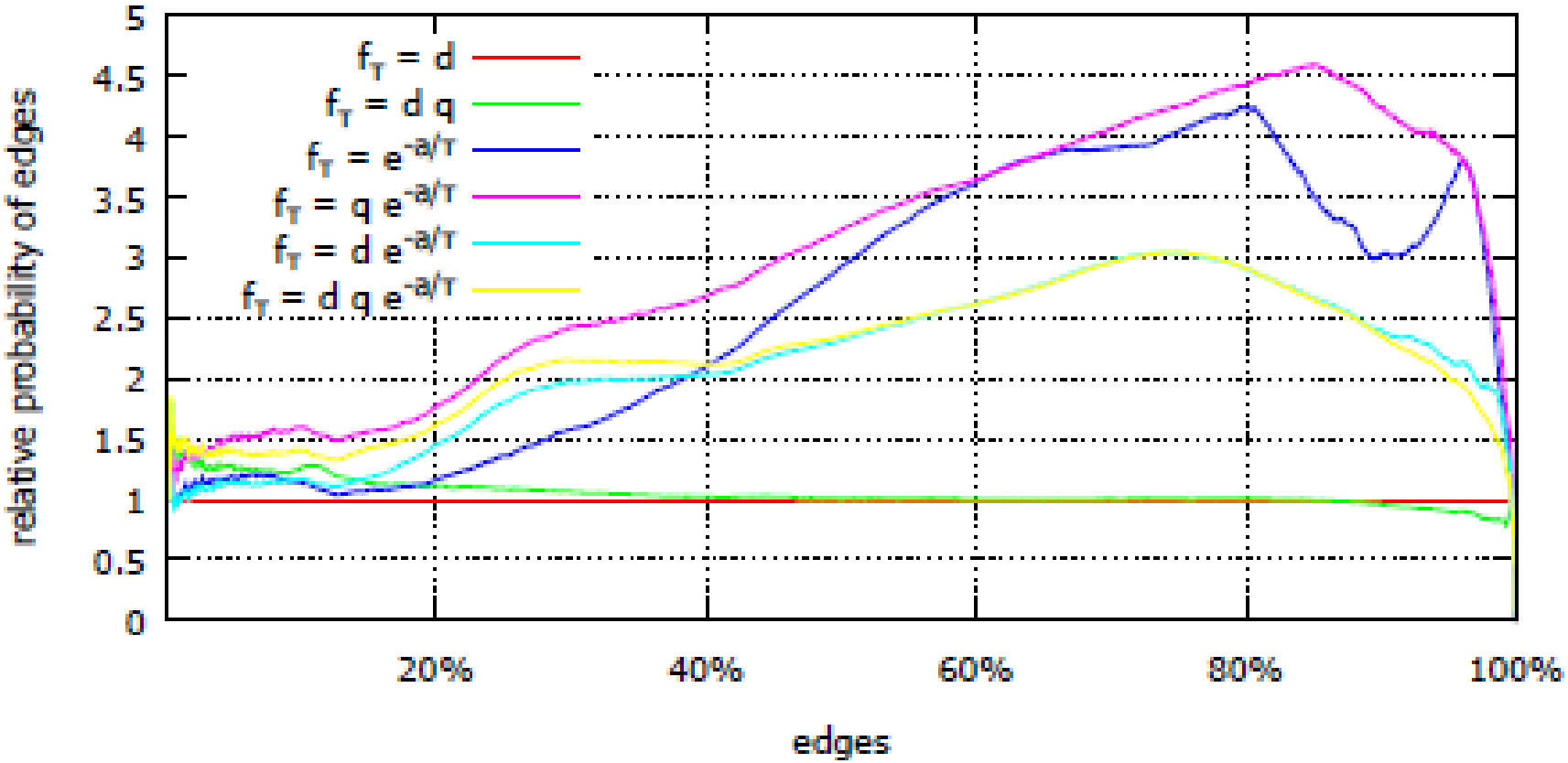
  and ζ_i are distributed according to the power-law
- In this case:
  - The graph is characterized by scale-free degree distribution
  - $e(T)$ is the exponentially decreasing function

Freshness model



Freshness model

Maximum Likelihood estimation results



Copy model

General idea

- Another idea to describe the Rich-Get-Richer phenomenon (scale-free properties)
- We will assume simply that people have a tendency to copy the decisions of people who act before them.
- Based on this idea, here is a simple model for the creation of links among Web pages
  - Pages are created in order, and named $1, 2, 3, \ldots, N$.
  - When page j is created, it produces a link to an earlier Web page according to the following probabilistic rule (which is controlled by a single number α between 0 and 1):
    - ⋆ With probability α, page j chooses a page i uniformly at random from among all earlier pages, and creates a link to this page i.
    - ⋆ With probability $1 - \alpha$, page j instead chooses a page i uniformly at random from among all earlier pages, and creates a link to the page that i points to
    - ⋆ This describes the creation of a single link from page j; one can repeat this process to create multiple, independently generated links from page j.

Copy model

General idea

- The main result about this model is that if we run it for many pages, the fraction of pages with k in-links will be distributed approximately according to a power law $1/k^{\gamma}$, where the value of the exponent γ depends on the choice of α
- This dependence goes in an intuitively natural direction: as α gets smaller, so that copying becomes more frequent, the exponent γ gets smaller as well, making one more likely to see extremely popular pages.
- the copying mechanism is really an implementation of the following "rich-get-richer" dynamics: when you copy the decision of a random earlier page, the probability that you end up linking to some page l is directly proportional to the total number of pages that currently link to l.

Copy model

Formal description

- G_0 is some fixed graph with t_0 nodes. Out-degrees of all nodes not less than d
- Assume that we already have G_t. Let's describe the process of G_{t+1} formation
- At time step $t+1$ we add one new node
- After that we successively add d edges $e_i,\ i = 1, \ldots, d$ going from the new node

Copy model

Formal description

- The process of links formation is as follows
  - We (uniformly) choose one random node p from the existing nodes in G_t. It's out-degree is at least d. Let $p_1, \ldots, p_d$ are random different node p neighbors
  - For each i we choose the end of the edge e_i as follows. With probability α we select random node from G_t. With probability $1 - \alpha$ we connect the new node with p_i.

Copy model

Properties

- Let $N_{t,r}$ be the number of nodes with in-degree r in the random graph on t nodes
- **Theorem.** Let $d = 1,\ r > 0$. Let $P_r = \lim \frac{\mathbb{E}N_{t,r}}{t}$.
  - P_r is defined and correct
  - The following estimate for P_r holds

$$P_r = P_0 \prod_{i=1}^{r} \frac{1 + \frac{\alpha}{i(1-\alpha)}}{1 + \frac{2}{i(1-\alpha)}} \longrightarrow P_r = \Theta\left(r^{-\frac{2-\alpha}{1-\alpha}}\right)$$

- **Theorem.** Let d be some fixed natural number, $\alpha \in (0,1)$, $t \to \infty$, $i < \ln t$. Then there exists $c = c(d)$:

$$P(\#(K_{i,d}, G_t) \geq cte^{-i}) \to 1,\ t \to \infty$$

There are a lot of nontrivial, big communities!

Copy model

Example

- $d = 10$
- $i = \ln t/2$
- With high probability the number of communities, having size i and citing 10 sites (the representation of the community members' interests), is about $\sqrt{t}$
- If $t = 10^10 \longrightarrow \ln t/2 \approx 11$, $\sqrt{t} = 100000$. In other words, we expect appearance of 100000 complete bipartite graphs on 10 and 11 nodes.

Holme–Kim model

- The main drawback of the studied previously preferential attachment models is an unrealistic behavior of the clustering coefficient.
- In fact, for all discussed models the clustering coefficient tends to zero as a graph grows
- In many real-world networks the clustering coefficient is approximately a constant
- A model with an asymptotically constant average local clustering coefficient was proposed by Holme and Kim.
- *Holme, P., Kim, B.J.: Growing scale-free networks with tunable clustering. Phys. Rev. E 65(2), 026107 (2002)*
- The idea is to mix preferential attachment steps with steps of triad formation.

Holme–Kim model

Steps of the graph creation:

(i) Initial condition: To start with, the network consists of m_0 vertices and no edges.

(ii) Growth: One vertex v with m edges is added at every time step. Time t is identified as the number of time steps.

(iii) Preferential attachment (PA): Each edge of v is then attached to an existing vertex with the probability proportional to its degree

(iv) **Triad formation (TF)**: If an edge between v and w was added in the previous PA step, then add one more edge from v to a randomly chosen neighbor of w. If there remains no pair to connect, i.e., if all neighbors of w were already connected to v, do a PA step instead.

When a vertex v with m edges is added to the existing network, we first perform one PA step, and then perform a TF step with the probability P_t or a PA step with the probability $1 - P_t$.

The average number m_t of the TF trials per added vertex is then given by $m_t = (m-1)P_t$, which we take as the control parameter in our model

Holme–Kim model

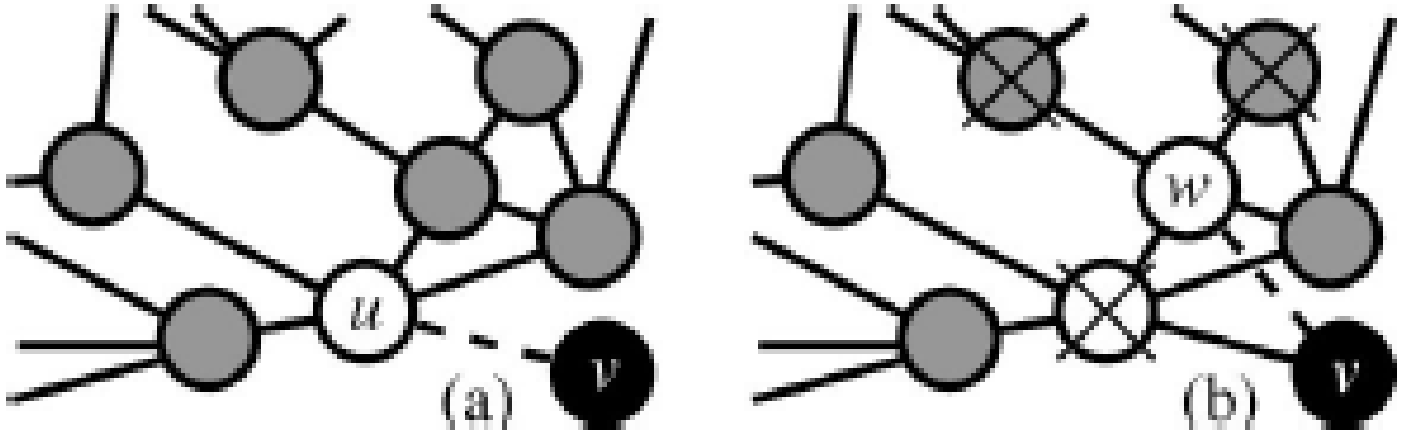


FIG. 1. *Preferential attachment and triad formation.* In the preferential attachment step (a) the new vertex v chooses a vertex u to attach to with a probability proportional to its degree. In the triad formation step (b) the new vertex v chooses a vertex w in the neighborhood of the one linked to in the previous preferential attachment step. $\times$ symbolizes "not allowed to attach to" (either since no triad would be formed, or that an edge already exists).

Holme–Kim model

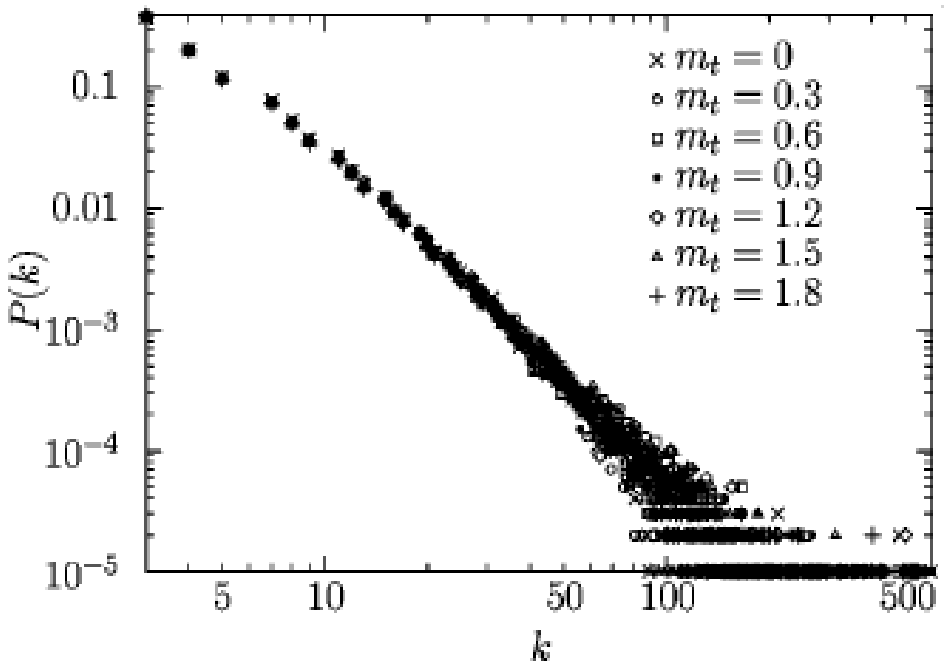


FIG. 2. Degree distribution for the scale-free network model with tunable clustering with parameter values $m = m_0 = 3$, $N = 10^5$ at various values of m_t: At any value of m_t, which determines the average number of triad formations, $P(k)$ exhibits a power-law behavior like the BA model corresponding to $m_t = 0$.

The degree distribution in this model obeys the power-law with the fixed parameter close to 3, which does not suit most real networks.

Holme–Kim model

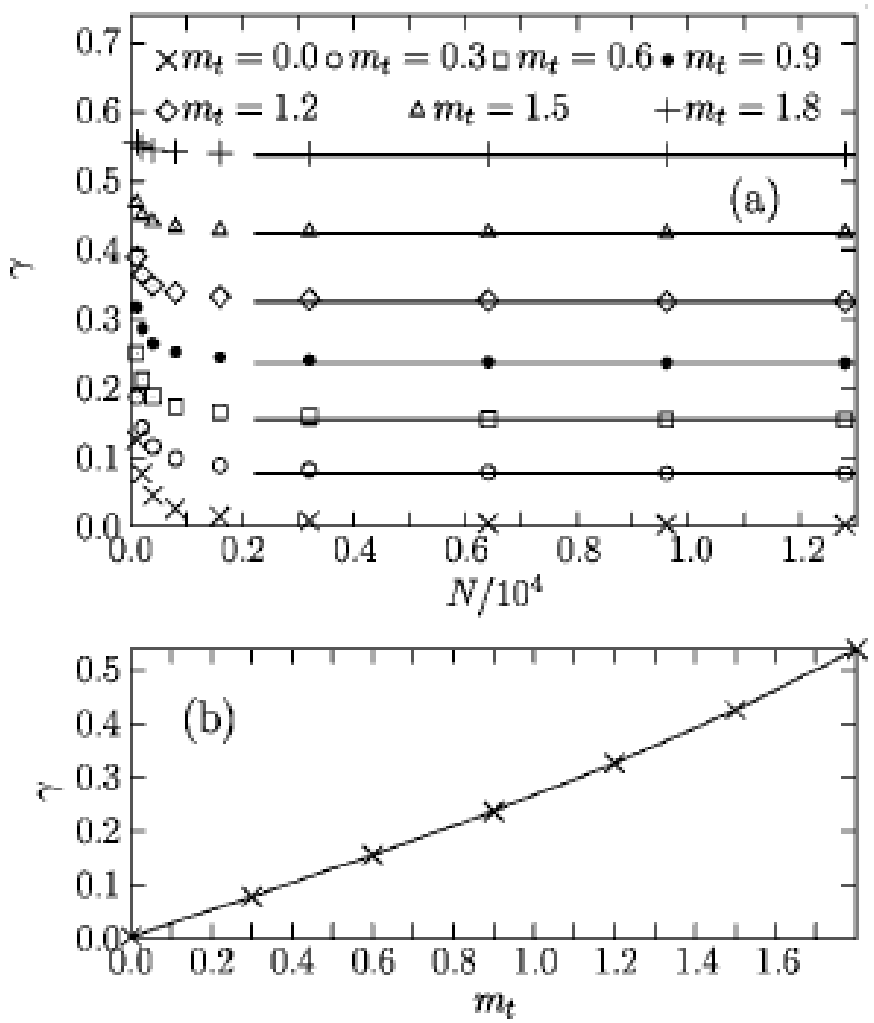


FIG. 3. (a) Clustering coefficient γ vs the network size N at various values of the average number m_t of triads per time step. Straight lines show asymptotic values of γ at each m_t. For $m_t \neq 0$, γ approaches a nonzero value as N is increased. (b) $\gamma(N \to \infty)$ vs m_t: The clustering coefficient can be varied systematically by changing m_t.