Hypergraph Model with Preferential Attachment for Scientific Collaborations

Hohyun Jung[*] Sung-Ho Kim[†] Frederick Kin Hing Phoa[‡]

September 27, 2023

Abstract

A hypergraph is useful to express the relationship between two or more nodes. Real hypergraph data are typically weighted. We propose a weighted evolving hypergraph model that considers preferential attachment. The model allows variability on the two basic components of the evolving hypergraph: the number and the size of the hyperedges to be connected. Under the mild distributional conditions on the two varying quantities, we derive the exact degree distribution that asymptotically follows a power-law distribution. We find that the limiting power-law exponent is affected by the distribution of hyperedge sizes. The distribution of the number of hyperedges to be connected has a considerable impact on a small-degree range in which non-power-law behavior is frequently observed in real data. Moreover, we argue that the degree distribution of the model can be expressed as a mixture of the degree distributions with a fixed number of hyperedges to be connected. The validity and usefulness of the model are explained with interpretations via a simulation study and real data analysis.

Key words: Degree distribution, Hypergraph, Preferential attachment, Network growth

1 The Model

Let $H = (V, E, w)$ be a *weighted hypergraph* with a set V of nodes, a set E of hyperedges, and a weight function $w : E \to \mathbf{R}$ from the hyperedge set to the real numbers, where $w(e)$ is the *weight* of hyperedge $e \in E$. A hyperedge $e = \{v_1, \cdots, v_n\}$ is a subset of V, and we call $|e| = n$ the *size* of a hyperedge. A *degree* $d(v)$ of a node v is defined as the sum of the weights of hyperedges that contain v, that is, $d(v) = \sum_{e\in E: v\in e} w(e)$.

We propose an evolving hypergraph model as follows:

- Initialization: Let the initial hypergraph be $H_0 = (V_0, E_0, w_0)$.

[*] School of Mathematics, Statistics and Data Science, Sungshin Women's University, Seoul 02844, South Korea (hhjung@sungshin.ac.kr)

[†] Department of Mathematical Sciences, KAIST, Daejeon, South Korea (sung-ho.kim@kaist.edu)

[‡] Institute of Statistical Science, Academia Sinica, Taiwan (fredphoa@stat.sinica.edu.tw)

- Growth: At each time $t = 1, 2, \cdots$, hypergraph $H_t = (V_t, E_t, w_t)$ is constructed by adding a new node v_{new} with new M_t hyperedges which are evoked by v_{new}. For each try $q = 1, \cdots, M_t$, v_{new} selects $Y^*_{q,t} = \min(Y_{q,t}, N_{t-1})$ existing nodes and forms a hyperedge consisting of v_{new} and the $Y^*_{q,t}$ nodes.

- Preferential attachment: The new node v_{new} selects an existing node v_i with probability proportional to the node degree $d(v_i)$.

Let $N_t = |V_t|$. Then we have $N_t = N_0 + t$. Further, we denote by $S_t = \sum_{v \in V_t} d(v)$ the sum of all the node degrees of the hypergraph H_t. We assume that there is at least one hyperedge in the initial hypergraph H_0, i.e., $N_0 \geq 2$ and $S_0 \geq 2$.

Let $\mathcal{D}_M$ be the distribution of the number of hyperedges to be connected with a probability mass function $P_M(m),\ m = 1, 2, \cdots$. Let $\mathcal{D}_Y$ be the distribution of the number of the nodes that are selected for a new hyperedge, with a probability mass function $P_Y(y),\ y = 1, 2, \cdots$. Let $\mu_M = E[M]$, $\sigma^2_M = Var[M]$, $\mu_Y = E[Y]$, and $\sigma^2_Y = Var[Y]$, where $M \sim \mathcal{D}_M$ and $Y \sim \mathcal{D}_Y$. We independently assign distributions on the two varying quantities that $M_t \sim \mathcal{D}_M$ and $Y_{q,t} \sim \mathcal{D}_Y,\ t = 1, 2, \cdots$, $q = 1, \cdots, M_t$. Those distributions have integer values larger than or equal to 1, implying that any newly coming node tries to construct at least one hyperedge.

We assume the following for our hypergraph model:

(A1) $E[M^\beta] < \infty,\ E[Y^\gamma] < \infty$ for some finite positive values β and γ such that

$$\frac{5}{2} < \beta < \infty,\ \mu_Y + 1 < \gamma < \infty,\ \left(\beta - \frac{5}{2}\right)(\gamma - 1) > \frac{3}{2}.$$

(A2) There exist an integer $t_0 \geq 0$ and a positive constant δ, $1/\gamma < \delta < 1/(\mu_Y + 1)$, satisfying

$$\max_{v \in V_t} d(v) \leq \frac{S_t}{N_t^\delta},\ \text{for } t = t_0, t_0 + 1, \cdots . \tag{1}$$

2 Degree Distribution

The *degree distribution* of a hypergraph is defined by the fraction of nodes in the hypergraph. Let $N_{k,t}$ be the number of nodes with degree k in H_t. Then we can express the degree distribution of H_t by $P_t(k) = N_{k,t}/N_t$. For the random generative process that we proposed in Section 2, we want to express its degree distribution at time t using $E[N_{k,t}]/N_t$. We say that the *steady-state degree distribution* of a hypergraph exists if $\lim_{t\to\infty}(E[N_{k,t}]/N_t)$ exists, and we write it as

$$P_\infty(k) = \lim_{t\to\infty} \frac{E[N_{k,t}]}{N_t}, \quad k = 0, 1, \cdots . \tag{2}$$

We now state and prove the degree distribution of the proposed evolving hypergraph model.

Theorem 1. *(Degree Distribution for Hypergraph) Assume (A1) and (A2). Then the steady-state degree distribution of a hypergraph exists and is given by*

$$P_\infty(k) = \frac{(1+1/\mu_Y)\,\Gamma(k)}{\Gamma(k+2+1/\mu_Y)} E\left[\frac{\Gamma\left(M+1+1/\mu_Y\right)}{\Gamma(M)} \mathbb{1}(M \leq k)\right],\ k = 1, 2, \cdots, \tag{3}$$

where the expectation is made over M.

Theorem 2. *Assume (A1) and (A2). Then the steady-state degree distribution asymptotically follows a power-law distribution with exponent $2 + 1/\mu_Y$. Specifically,*

$$P_\infty(k) \approx_k (1+1/\mu_Y)\, E\left[\frac{\Gamma\left(M+1+1/\mu_Y\right)}{\Gamma(M)}\right] k^{-(2+1/\mu_Y)}, \tag{4}$$

where the sign $\approx_k$ means that the ratio of two quantities on both sides of the sign tends to 1 as k tends to infinity.

Theorem 3. *Assume (A1) and (A2). Suppose that M has a finite support, i.e., $M \leq M_{max}$ for some integer M_{max}. Then the steady-state degree distribution of the model is a finite mixture of the distributions $P_{\infty,m}(k)$ of the model with constant m for $m = 1, \cdots, M_{max}$, given by*

$$P_\infty(k) = \sum_{m=1}^{M_{max}} P_M(m) P_{\infty,m}(k). \tag{5}$$

3 Numerical Simulation

In this section, we perform extensive experiments as a way of validation of the result in Theorem 1 under various distributional settings of M and Y. We use Python for simulation, and the numpy.random.choice[1] function of NumPy version 1.16.4 is employed for the weighted choice of nodes. The initial hypergraph H_0 is set as the hypergraph consisting of ten nodes $V_0 = \{1, \cdots, 10\}$ and two hyperedges $E_0 = \{\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, 10\}\}$ of size 5 with unit weights. Then we add $10,000$ nodes at time points $t = 1, \cdots, 10,000$ according to the process in Section 1. We focus on the following two key features of the derived limiting degree distribution: the impact of $\mathcal{D}_Y$ on the limiting power-law behavior and the impact of $\mathcal{D}_M$ on the non-power-law behavior in a range of small degrees.

According to Theorem 1, the degree distribution is affected by $\mathcal{D}_Y$ only through μ_Y. Also, Theorem 2 shows that the limiting power-law exponent is $2 + 1/\mu_Y$. To check these properties, we fix $M = 2$, and each synthetic data is generated with (1a) $Y = 2$, (1b) $Y = 1, 2, 3$ with $P_Y(1) = 0.25$, $P_Y(2) = 0.5$, $P_Y(3) = 0.25$, and (1c) $Y \sim 1 + Poisson(1)$. For all the three cases, we have $\mu_Y = 2$.

[1] https://docs.scipy.org/doc/numpy-1.16.0/reference/generated/numpy.random.choice.html

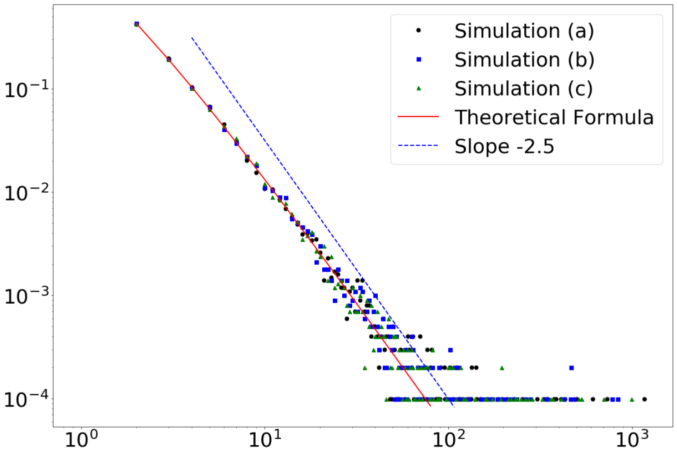


Figure 1: The degree distribution $P(k)$ for the generated hypergraphs and the theoretical steady-state degree distribution $P_\infty(k)$ in Theorem 1. Hypergraphs are generated with $M = 2$ and (a) $Y = 2$, (b) $Y = 1, 2, 3$ with $P_Y(1) = 0.25$, $P_Y(2) = 0.5$, $P_Y(3) = 0.25$, and (c) $Y \sim 1 + Poisson(1)$.

Figure 1 shows the degree distribution of the generated hypergraphs. Although the distributions of Y are different, the degree distributions show almost the same shape. Moreover, the simulation results agree well with the theoretical result. We draw the line with slope $-(2 + 1/\mu_Y) = -2.5$, and it indicates that the simulation data tends to follow the limiting ($k \to \infty$) slope -2.5.

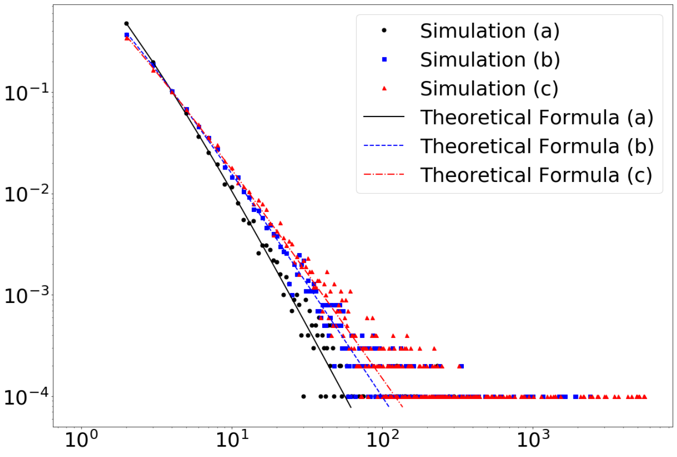


Figure 2: The degree distribution $P(k)$ for the generated hypergraphs and the theoretical steady-state degree distribution $P_\infty(k)$ in Theorem 1. Hypergraphs are generated with $M = 2$ and (a) $Y \sim 1 + Poisson(0.25)$, (b) $Y \sim 1 + Poisson(3)$, and (c) $Y \sim 1 + Poisson(9)$.

We generate data with the Poisson distribution on Y, (2a) $Y \sim 1 + Poisson(0.25)$, (2b) $Y \sim 1 + Poisson(3)$, and (2c) $Y \sim 1 + Poisson(9)$. Again, $M = 2$ is fixed. The average values of Y are different over data, 1.25, 4, and 10 for (2a), (2b), and (2c), respectively. In Figure 2, we plot the degree distributions of the simulated data. The result indicates that as the quantity μ_Y increases, the power-law exponent of the degree distribution decreases. It also tends to obey the theoretical result well.

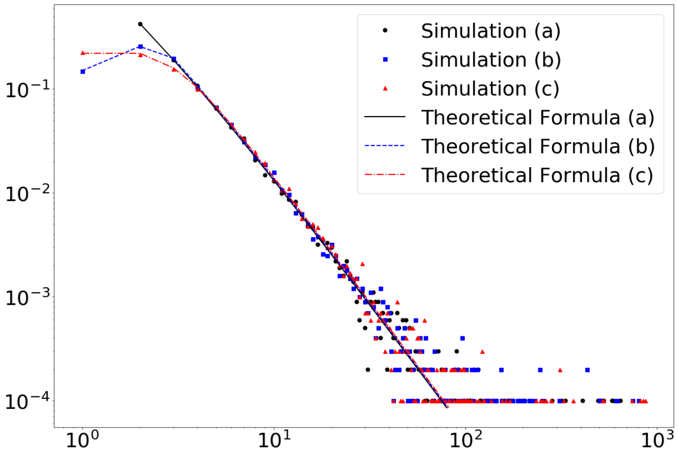


Figure 3: The degree distribution $P(k)$ for the generated hypergraphs and the theoretical steady-state degree distribution $P_\infty(k)$ in Theorem 1. Hypergraphs are generated with $Y = 2$ and (a) $M = 2$, (b) $M = 1, 2, 3$ with $P_M(1) = 0.25$, $P_M(2) = 0.5$, $P_M(3) = 0.25$, and (c) $M \sim 1 + Poisson(1)$.

We now fix $Y = 2$ and vary M as (3a) $M = 2$, (3b) $M = 1, 2, 3$ with $P_M(1) = 0.25$, $P_M(2) = 0.5$, $P_M(3) = 0.25$, and (3c) $M \sim 1 + Poisson(1)$. Figure 3 shows the degree distribution of generated hypergraphs. The assumption of varying M has a significant influence on a small degree range. The degree distribution of the case (3a) seems straight from $k = 2$. The case (3b) shows a straight line shape from $k = 3$, and we can observe that it is curved for $k \leq 3$. In the case of (3c), the degree distribution is getting straightened as k increases. This result suggests that the degree distribution can be flexibly modified in a small degree range according to the distribution of M.

4 Real Data Analysis

We here analyze scientific collaboration hypergraphs obtained from the Web of Science, which provides all the published scientific articles in the world. The eight fields are considered in our analysis: Biotechnology & Applied Microbiology, Computer Science, Electrical & Electronic Engineering, Genetics & Heredity, Management, Physical Chemistry, Sociology, and Statistics & Probability. These hypergraphs are of the scientific papers published from 2007 to 2016. The nodes represent authors and hyperedges co-authorships. A hyperedge consists of the set of authors of a paper. In order to investigate the scientific collaboration, all single-author papers are excluded from the analysis.

Table 1: Summary statistics and power-law fitting results of the scientific collaboration data. Eight fields are considered. We present the number of nodes and hyperedges in 2016, the average and standard deviation of the degree distribution in 2016, and the average and standard deviation of calculated M and Y. We also present the estimated k_{min} and α according to BIC.

Field	Number of		Degree		BIC		$\mathcal{D}_M$		$\mathcal{D}_Y$	
	Nodes (N)	Hyperedges	Avg.	Std.	$\hat{k}_{min}$	$\hat{\alpha}$	Avg.	Std.	Avg.	Std.
Biotechnology & A. M.	729,478	256,975	2.01	3.67	12	2.96	1.52	1.30	4.49	3.37
Computer Science	528,267	325,396	2.37	4.71	14	2.95	1.61	1.42	2.49	1.92
Electrical & Electronic E.	591,093	364,216	2.74	6.00	22	2.98	1.80	1.87	2.94	2.17
Genetics & Heredity	645,771	207,195	2.32	4.56	15	2.92	1.62	1.67	6.00	8.25
Management	97,229	61,423	1.97	2.55	10	3.60	1.44	1.03	1.85	1.15
Physical Chemistry	727,213	450,109	3.27	7.70	24	2.97	2.02	2.28	3.82	2.40
Sociology	42,517	21,697	1.50	1.41	6	3.63	1.27	0.73	1.72	1.17
Statistics & Probability	92,397	59,023	2.24	3.83	11	3.05	1.56	1.37	2.01	1.62

In order to investigate the impact of $\mathcal{D}_M$ and $\mathcal{D}_Y$ on the observed degree distributions, we fit the Zipf distribution to the degree distribution in 2016 for each field. The Zipf distribution, denoted by $Zipf(\alpha, k_{min})$, is a discrete form of the power-law distribution with a probability mass function

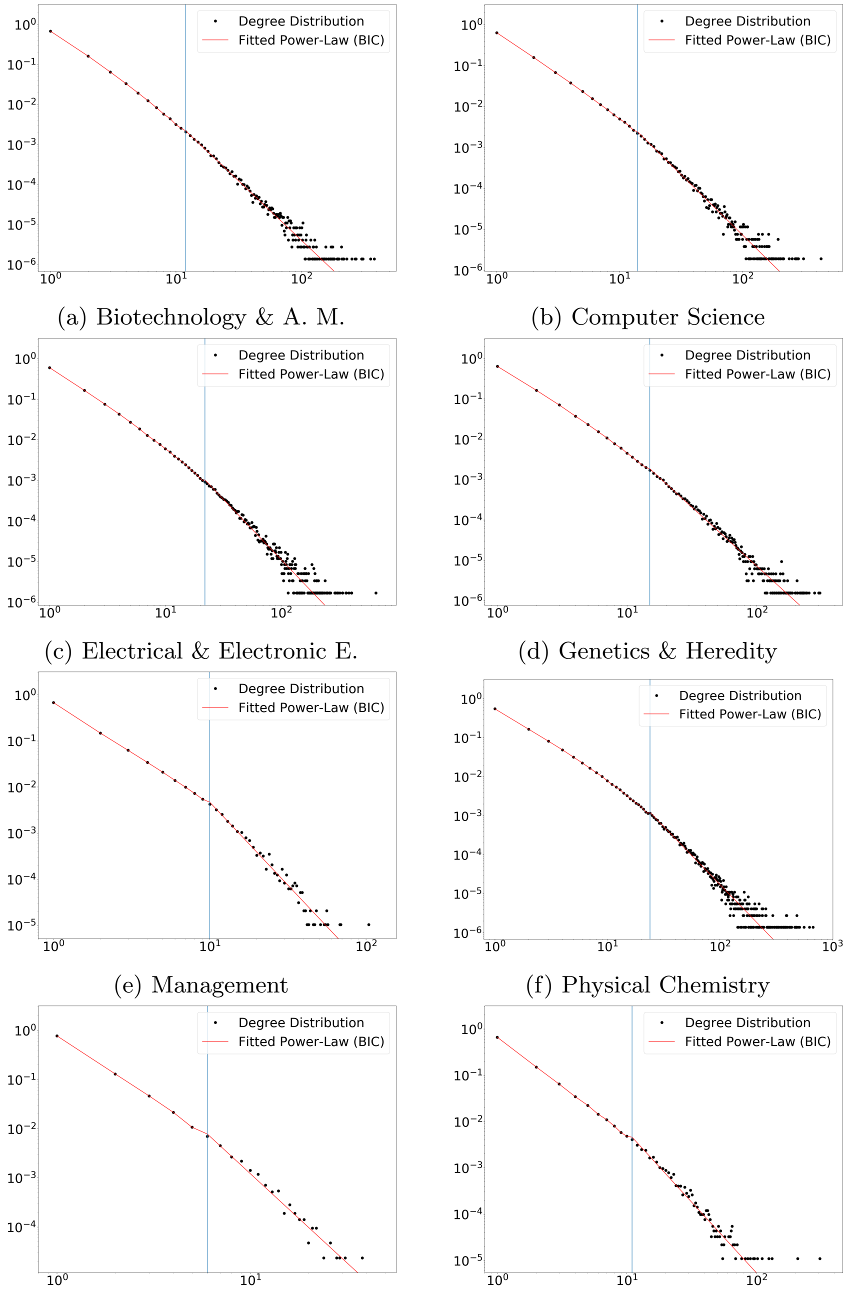
$$f(k|\alpha, k_{min}) = \frac{1}{\zeta(\alpha, k_{min})} k^{-\alpha}, \quad k = k_{min}, k_{min}+1, \cdots, \tag{6}$$

where $\zeta(\alpha, k_{min}) = \sum_{k=k_{min}}^{\infty} k^{-\alpha}$ is a Hurwitz zeta function. Note that $f(k|\alpha, k_{min}) \propto k^{-\alpha}$. Let $k_i = d(v_i)$ be the degree of author v_i, $i = 1, 2, \cdots, N$, where N is the number of authors in the system. Then we choose the power-law exponent α and the smallest value k_{min} of the range of the node degrees over which the degree distribution is in a power-law shape. This is made by applying the approach of Handcock and Jones (2004) [1]. We estimate the degree distribution for the data given by

$$p(k|\pi, \alpha, k_{min}) = \begin{cases} \pi_k & \text{if } \ k = 1, \cdots, k_{min}-1, \\ \left(1 - \sum_{k'=1}^{k_{min}-1} \pi_{k'}\right) f(k|\alpha, k_{min}) & \text{if } \ k = k_{min}, k_{min}+1, \cdots, \end{cases}$$

where $\pi = (\pi_1, \cdots, \pi_{k_{min}-1})$ and $f(k|\alpha, k_{min})$ is as in Eq. (6). Once k_{min} is chosen, the parameters α and π_k, $k = 1, 2, \cdots, k_{min}-1$ are determined by maximizing the likelihood function $\mathcal{L}(\pi, \alpha|k_1, \cdots, k_N) = \prod_{i=1}^{N} p(k_i|\pi, \alpha, k_{min})$.

Then, how can we determine k_{min}? The estimates of α and the fitting line might vary according to chosen k_{min}. The importance and methodology of choosing k_{min} are thoroughly discussed in Clauset et al. (2009) [2]. In our study, we choose k_{min} that minimizes $BIC = -2\ln\mathcal{L}(\hat{\pi}, \hat{\alpha}|k_1, \cdots, k_N) + (\ln N)k_{min}$, where $\hat{\pi}$ and $\hat{\alpha}$ are the maximum likelihood estimates corresponding to k_{min}. Table 1



(a) Biotechnology & A. M.

(b) Computer Science

(c) Electrical & Electronic E.

(d) Genetics & Heredity

(e) Management

(f) Physical Chemistry

(g) Sociology

(h) Statistics & Probability

Figure 4: Degree distribution and fitted power-law curves for the eight fields of scientific collaboration hypergraphs according to BIC. The estimated k_{min} values are depicted by blue vertical lines.

presents the estimates of k_{min} and α according to BIC. The degree distribution and fitted power-law curves are shown in Figure 4. We can see that k_{min} (depicted by the vertical line) is reasonably determined for all the eight fields.

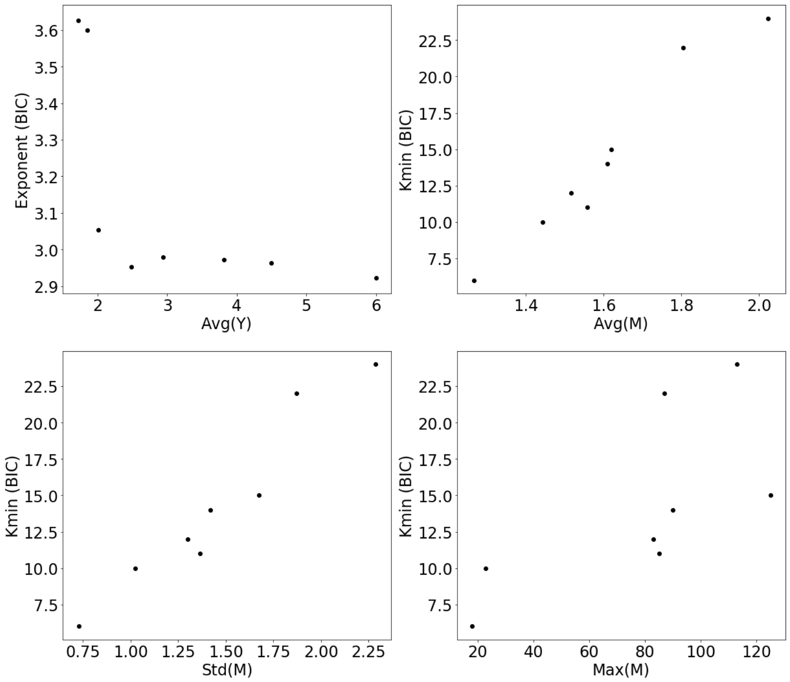


Figure 5: The relationships between the expectation of Y and $\hat{\alpha}$ (upper left), the expectation of M and $\hat{k}_{min}$ (upper right), the standard deviation of M and $\hat{k}_{min}$ (lower left), and the maximum of M and $\hat{k}_{min}$ (lower right) according to BIC.

In Figure 5, we plot some relationships among $\hat{\alpha}$, $\hat{k}_{min}$, $\hat{\mathcal{D}}_M$, and $\hat{\mathcal{D}}_Y$. We can observe that (i) $\hat{\alpha}$ is inversely proportional to the average of Y, and (ii) $\hat{k}_{min}$ is proportional to the average, standard deviation, and maximum of M.

5 Concluding Remarks

In this paper, we proposed a novel weighted hypergraph model considering the preferential attachment. We believe that the proposed model is productive since we allowed variabilities on the two primary constituents of the evolving hypergraph: the number and size of the hyperedges to be connected. We have shown that the exact degree distribution exists under the mild conditions of M and Y. Surprisingly, the degree distribution has an asymptotic power-law behavior, and the limiting power-law exponent is only affected by the average of Y. The special cases of fixed M and Y were studied in this work, where their degree distributions turned out to have a simpler form than the general case of the model. We also argued that the case of varying M can be expressed as a mixture distribution of fixed M cases. The deviation of a power-law behavior in a small degree range was able to be explained by the varying M assumption. The theoretical degree distribution was evidenced by the extensive simulation studies.

References

[1] M. S. Handcock and J. H. Jones, "Likelihood-based inference for stochastic models of sexual network formation," *Theoretical Population Biology*, vol. 65, no. 4, pp. 413–422, 2004.

[2] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.