www.sciencemag.org/cgi/content/full/325/5948/1696/DC1

Supporting Online Material for

On Universality in Human Correspondence Activity

R. Dean Malmgren,* Daniel B. Stouffer, Andriana S. L. O. Campanharo, Luís A. Nunes Amaral*

*To whom correspondence should be addressed. E-mail: dean.malmgren@u.northwestern.edu (R.D.M.); amaral@northwestern.edu (L.A.N.A.)

Published 25 September 2009, *Science* **325**, 1696 (2009)
DOI: 10.1126/science.1174562

This PDF file includes:

Methods

Consider a cascading Poisson process with parameters $\boldsymbol{\theta} = \{\rho, \xi\}$. The inter-event time distribution is given by

$$p\left(\tau|\boldsymbol{\theta}\right) \;=\; \left\{ \begin{array}{ll} \xi & \tau = 0 \\ (1-\xi)\rho e^{-\rho\tau} & \tau > 0 \end{array} \right. , \tag{S1}$$

and the probability of observing $N_{T_\star}$ events during a time interval of duration $T_\star$ can be written as

$$p\left(N_{T_\star}|\boldsymbol{\theta}\right) = \begin{cases} e^{-\rho T_\star} & N_{T_\star} = 0 \\ e^{-\rho T_\star} Q(N_{T_\star} - 1; \boldsymbol{\theta}, T_\star) & N_{T_\star} > 0 \end{cases} , \tag{S2}$$

where the polynomial

$$Q(N; \boldsymbol{\theta}, T_\star) \;=\; (1-\xi)\rho T_\star \sum_{n=0}^{N} \binom{N}{n} \frac{\xi^n \left[(1-\xi)\rho T_\star\right]^{N-n}}{(n+1)!} \tag{S3}$$

accounts for the various ways that the $N = N_{T_\star} - 1$ events during the time interval of duration $T_\star$ time units can be grouped into cascades of activity. The censored likelihood function is given by

$$\mathcal{L}(\boldsymbol{\theta}) \;=\; \prod_{k=0}^{T/T_\star - 1} \Pr(N_{T_\star,k}|\boldsymbol{\theta}) \, , \tag{S4}$$

where T is the duration of the time segment, $T_\star = 1$ day, and $N_{T_\star,k}$ is the number of events that occur on day k. The derivation for these quantities can be found in Sec. S4.

S1 Preprocessing the data

The empirical data consists of letters sent or received by 16 writers, performers, politicians, and scientists (*S1,S2,S3,S4,S5,S6,S7,S8,S9,S10,S11,S12,S13*). In our study, we focused on the letters that these individuals sent. There are a number of issues with the data that mandate preprocessing. For example, according to the records, almost of 19% of Ernest Hemingway's letters have either unspecified or ambiguous authorship dates (e.g., "`Aug 1945`", "`1946/47`", "`Early 1950`", "`1960?`" or "`Fall 1960`") (*S12*). We remove all letters in the data for which the precise date is unknown.

Additionally, the letter correspondence records are aggregated from a variety of sources. Some of the letters are carbon copies that were saved by the original author. Other letters are collected from the original recipients of the letters and returned to the database. Having a complete letter correspondence record for a particular individual, therefore, either relies on (*i*) an individual to retain a copy of each letter, (*ii*) all recipients of an individual's letters to retain a copy of their letters, or (*iii*) same fortuitous combination of (*i*) and (*ii*). We have confirmed that our results are robust with regard to these anomalies in the data collection method (Sec. S2).

In the case of Albert Einstein, there is one more challenge: several letters appear to be duplicates arising from the fact that the data is collected from different sources. To illustrate the difficulty in identifying duplicate entries, consider the two letters sent on September 25, 1907 to Joseph Stark and Johannes Stark, both in Griefswald, Germany. According to the database, the letter to Joseph Stark is a typed transcript of a letter (denoted TTRL in the database) and the letter to Johannes Stark is a xerox copy of a handwritten and signed letter (ALSX). While it is conceivable that Einstein sent a letter to Joseph Stark and another letter to Nobel Laureate Johannes Stark, we think it is more likely that the letter addressed to Joseph Stark is actually a draft of the same letter addressed to Johannes Stark, of which the database has a xerox copy. As this example illustrates, we can not simply use the designation ALSX or TTRL to detect duplicate letters, we must also use their names. To overcome this difficulty we use a dynamic programming text matching algorithm (*S14*) to semi-automatically detect if letters are duplicates; that is, exceptionally different recipients are automatically detected and we manually curate recipients who have marginally similar names. This procedure excludes another 651 letters written by Einstein.

In summary, we exclude from our analysis letters with uncertain dates and duplicate letters. The results of our preprocessing procedure are summarized in Tbl. S1.

S2 Robustness of results with regard to data collection method

In e-mail correspondence, it is relatively trivial to collect correspondence activity; e-mail correspondence can easily be extracted from the log files of an e-mail server. Collecting letter correspondence data is not so simple. Unlike e-mail servers, the postal service does not archive all written communications, so it is not possible to simply query the postal service for all correspondence written by a particular individual. Instead, collecting correspondence records relies on the letter authors or recipients to save letters and then return them to a centralized database. We examine two possible scenarios in which the limitations of this data collection method could potentially distort our results.

In the first scenario, only a fraction of the total volume of letters originally sent by an individual are actually saved and compiled in a centralized database. This will almost certainly be an issue for almost every individual, since it is highly unlikely that every letter is saved by either the author or the recipient. To test whether such an artifact of the data collection method might affect our conclusions about the validity of the cascading Poisson process, we randomly select a fraction of the letters that Schoenberg sent. Although the resulting parameter estimates predictably decrease as fewer and fewer letters are retained during our analysis (Fig. S1), our Monte Carlo hypothesis testing results confirm that this artifact of data collection does not affect our conclusion that a cascading Poisson process is consistent with the empirical data (Tbl. S2). Importantly, although we have simulated the loss rate to be uniform over Schoenberg's lifetime, a non-uniform loss rate will not affect our results provided that the loss rate during each time segment is approximately the same.

In the second scenario, only certain individuals might save letters and return them to the centralized database. In the most extreme case, only one individual, perhaps a close friend or family member, might save their correspondence. To test whether such an artifact of the data collection method might affect our conclusions about the validity of the cascading Poisson process, we consider Charles Darwin's correspondence to his close friend, the well-known botanist J.D. Hooker. Darwin sent Hooker 797 letters between 1844 and 1882. After segmenting this time series to account for non-stationarities in the communication from Darwin to Hooker, we obtain 31 time segments. Monte Carlo hypothesis testing rejects 1 of the 31 time segments, which is within the 95% confidence interval $[0, 4]$ of the corresponding binomial model, indicating that a cascading Poisson process is still consistent with the data in spite of the bias in the sampling of Darwin's correspondence. Obviously, the resulting parameter estimates for Darwin's correspondence to Hooker are significantly different than the parameter estimates from the correspondence to all recipients (Fig. S2). In particular, we note that cascades of activity cease to be important since it is highly unlikely that someone would send more than one letter to an individual in the same day.

These results demonstrate that these artifacts of letter correspondence data collection do not obfuscate our primary claim that a cascading Poisson process is consistent with the letter correspondence patterns of the individuals under consideration, regardless of whether the correspondence records are sampled uniformly at random or whether the correspondence records are sampled non-uniformly.

S3 Other candidate models

We have conducted the same Monte Carlo hypothesis testing procedure for three other candidate models, the results of which are summarized in Tbl. S3 and Fig. S3. In the limit that cascades of activity and weekly periodicities are irrelevant, a homogeneous Poisson process may be a reasonable candidate model for letter correspondence. This model has a single parameter—the rate of sending letters ρ_i—that is readily estimated using maximum likelihood (Sec. S4) during each stationary time segment. *This model is rejected for 7 individuals.*

In the limit that cascades of activity are irrelevant but weekly cycles of activity are important, a non-homogeneous Poisson process may be a reasonable candidate model for letter correspondence. Here, we assume that the non-homogeneous Poisson process is periodic on the weekly scale, so this model has seven parameters—the rate of sending letters $\rho_{i,t}$ during each day of the week t—that are readily estimated using maximum likelihood for each stationary time segment. *This model is rejected for 6 individuals.*

If, as in the case of e-mail correspondence, cascades of activity and weekly cycles are important, a cascading non-homogeneous Poisson process may be a reasonable candidate model for letter correspondence. Here, we assume that the cascading non-homogeneous Poisson process is periodic on the weekly scale, so this model has eight parameters—the rate of sending letter $\rho_{i,t}$ during each day of the week t as well as the probability ξ_i of sending additional letters during cascades of activity—that are readily estimated using maximum likelihood for each stationary time segment. Like the cascading Poisson process presented in the manuscript, *this model can not be rejected for any individual*, however the increased complexity of the cascading non-homogeneous Poisson process is unwarranted since the simpler, two-parameter model is equally descriptive of letter correspondence.

These results illustrate a few interesting features of models that are necessary for describing letter correspondence patterns. First, based on the success of the models that include cascading versus those that do not, we infer that cascades of activity are an essential element for describing letter correspondence. Importantly, cascades of activity are also essential for describing e-mail correspondence patterns (*S15*). Second, since the models that include weekly periodicities have no greater explanatory power than the models that do not include weekly periodicities, we conclude that weekly patterns of activity are not an essential element for describing letter correspondence. This suggests that letter correspondence does not appear to have the same dependence on the weekly work cycle as e-mail correspondence.

S4 Analytical results

Before we derive the likelihood function for a cascading Poisson process where the data are censored, it is illustrative to first pedagogically demonstrate how to derive the likelihood function for a homogeneous Poisson process in the absence and presence of censoring and then for a cascading Poisson process in the absence and presence of censoring. In our derivations of the parameter estimates $\boldsymbol{\theta}$ for these models, we consider a time series $\{t_1, t_2, \ldots, t_N\}$ of N ordered events occurring within time segment $[0, T)$. For clarity, we omit the index i which was used throughout the manuscript to denote the parameters $\boldsymbol{\theta}_i$ during time segment i.

Homogeneous Poisson process. A homogeneous Poisson process with parameters $\boldsymbol{\theta} = \{\rho\}$ predicts that, during an infinitesimal time window of duration dt, an event either occurs (denoted by $\bullet$) at time t with probability $\mathrm{Pr}_\bullet(t) = \rho dt$ or does not occur (denoted by $\circ$) at time t with probability $\mathrm{Pr}_\circ(t) = (1 - \rho dt)$. Note that for a homogeneous Poisson process the outcome at time t is independent of the outcome at time $t - dt$. Given an observed sequence of N ordered events $0 \leq t_1 \leq t_2 \leq \cdots \leq t_N < T$ during time segment $[0, T)$, the probability that this sequence was generated from a homogeneous Poisson process is given by

$$\begin{aligned}
\mathrm{Pr}(t_1, t_2, \ldots, t_N | \boldsymbol{\theta}) = & \left[\prod_{k=0/dt}^{t_1/dt-1} \mathrm{Pr}_\circ(kdt)\right] \mathrm{Pr}_\bullet(t_1) \left[\prod_{k=t_1/dt+1}^{t_2/dt-1} \mathrm{Pr}_\circ(kdt)\right] \mathrm{Pr}_\bullet(t_2) \cdots \\
& \mathrm{Pr}_\bullet(t_N) \left[\prod_{k=t_N/dt+1}^{T/dt-1} \mathrm{Pr}_\circ(kdt)\right] \\
= & (1 - \rho dt)^{(t_1 - 0)/dt} \rho dt (1 - \rho dt)^{(t_2 - t_1)/dt - 1} \rho dt \cdots \\
& \rho dt (1 - \rho dt)^{(T - t_N)/dt - 1}.
\end{aligned}$$

Note that

$$\begin{aligned}
\lim_{dt \to 0} (1 - \rho dt)^{\Delta t/dt - 1} &= \lim_{dt \to 0} \frac{(1 - \rho dt)^{\Delta t/dt}}{(1 - \rho dt)} \\
&= \frac{e^{-\rho dt(\Delta t/dt)}}{1} \\
&= e^{-\rho \Delta t}.
\end{aligned}$$

Using this result, we obtain the likelihood function for a homogeneous Poisson process in the limit that $dt \to 0$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \Pr(t_1, t_2, \ldots, t_N|\boldsymbol{\theta}) \\ &= e^{-\rho(t_1-0)}\rho dt e^{-\rho(t_2-t_1)}\rho dt \cdots \rho dt e^{-\rho(T-t_N)}, \\ &= (\rho dt)^N e^{-\rho T}. \end{aligned} \tag{S5}$$

By taking the derivative of $\log \mathcal{L}(\boldsymbol{\theta})$ with respect to the rate ρ, it is straightforward to see that the likelihood function is optimized with the best-estimate rate $\widehat{\rho} = N/T$ for a homogeneous Poisson process.

When the data are interval censored, as is the case of letter correspondence, our approach to estimating the parameters changes to reflect our uncertainty in the precise timing of events. For instance, suppose that our data have a resolution of $T_\star = 1$ day and that, on a particular day, we observe that $N_{T_\star}$ events occurred. Assuming that our data are generated by a homogeneous Poisson process, the probability that $N_{T_\star}$ events occurred during a time interval of duration $T_\star = 1$ day is given by marginalizing the likelihood $\Pr\big(t_1, t_2, \ldots, t_{N_{T_\star}}|\boldsymbol{\theta}\big)$ over all possible configurations of an ordered set of events $0 \leq t_1 \leq t_2 \leq \cdots \leq t_{N_{T_\star}} < T_\star$ occurring during this interval:

$$\begin{aligned}\Pr(N_{T_\star}|\boldsymbol{\theta}) &= \int\limits_0^{T_\star}\int\limits_{t_1}^{T_\star}\cdots\int\limits_{t_{N_{T_\star}-1}}^{T_\star} \rho^{N_{T_\star}} e^{-\rho T_\star} dt_{N_{T_\star}} \cdots dt_2 dt_1 \\ &= \rho^{N_{T_\star}} e^{-\rho T_\star}\int\limits_0^{T_\star}\int\limits_{t_1}^{T_\star}\cdots\int\limits_{t_{N_{T_\star}-1}}^{T_\star} dt_{N_{T_\star}} \cdots dt_2 dt_1 \\ &= \frac{(\rho T_\star)^{N_{T_\star}} e^{-\rho T_\star}}{N_{T_\star}!}, \end{aligned} \tag{S6}$$

resulting in the well-known Poisson distribution. Then, to estimate the parameters of a homogeneous Poisson process over the entire time segment $[0, T)$, we can account for the interval censoring during parameter estimation by writing down the probability of observing $N_{T_\star,k}$ events on each day k as

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{k=0}^{T/T_\star - 1} \Pr(N_{T_\star,k}|\boldsymbol{\theta}) \tag{S7}$$

where $T/T_\star$ is the number of days during time segment $[0, T)$. By taking the derivative of $\log \mathcal{L}(\boldsymbol{\theta})$ with respect to the rate ρ, we again find that the best-estimate rate $\widehat{\rho} = N/T$. Although this result is exactly the same for a homogeneous Poisson process regardless of whether the data is interval censored or not, the important distinction is that maximum likelihood parameter estimation in the interval censored case explicitly depends on $\Pr(N_{T_\star}|\boldsymbol{\theta})$. This fact is important to consider when deriving the censored likelihood function for the cascading Poisson process.

Cascading Poisson process. Recall that in our cascading Poisson process, cascades of events are initiated by a homogeneous Poisson process with rate ρ and that each additional event in the cascade occurs with probability $\xi \gg \rho dt$. A cascading Poisson process with parameters $\boldsymbol{\theta} = \{\rho, \xi\}$ therefore predicts that, during an infinitesimal time window of duration dt, an event either occurs (denoted by $\bullet$) or does not occur (denoted by $\circ$) depending on whether an event occurred at time $t - dt$: if an event did not occur at time $t - dt$, then an event occurs at time t with probability $\Pr_{\circ\bullet}(t) = \rho dt$ or does not occur with probability $\Pr_{\circ\circ}(t) = (1 - \rho dt)$; if an event did occur at time $t - dt$, then an event occurs at time t with probability $\Pr_{\circ\bullet}(t) = \xi$ or does not occur with probability $\Pr_{\circ\circ}(t) = (1 - \xi)$. Then, given a sequence of N ordered events, the probability that this sequence was generated from a cascading Poisson process during the time segment $[0, T)$ is given by

$$
\begin{aligned}
\Pr(t_1, t_2, \ldots, t_N|\boldsymbol{\theta}) = {} & \left[\prod_{k=0/dt}^{t_1/dt-1} \Pr_{\circ\circ}(kdt)\right] \Pr_{\circ\bullet}(t_1) \\
& \left\{\delta_{t_1+dt,t_2} \Pr_{\bullet\bullet}(t_2) + \right. \\
& \quad \left. (1-\delta_{t_1+dt,t_2}) \Pr_{\bullet\circ}(t_1+dt) \left[\prod_{k=t_1/dt+2}^{t_2/dt-1} \Pr_{\circ\circ}(kdt)\right] \Pr_{\circ\bullet}(t_2)\right\} \cdots \\
& \left\{\delta_{t_{N-1}+dt,t_N} \Pr_{\bullet\bullet}(t_N) + \right. \\
& \quad \left. \left(1-\delta_{t_{N-1}+dt,t_N}\right) \Pr_{\bullet\circ}(t_{N-1}+dt) \left[\prod_{k=t_{N-1}/dt+2}^{t_N/dt-1} \Pr_{\circ\circ}(kdt)\right] \Pr_{\circ\bullet}(t_N)\right\} \\
& \Pr_{\bullet\circ}(t_N+dt) \left[\prod_{k=t_N/dt+2}^{T/dt-1} \Pr_{\circ\circ}(kdt)\right]
\end{aligned}
$$

$$
\begin{aligned}
\Pr(t_1, t_2, \ldots, t_N|\boldsymbol{\theta}) = {} & \left[(1-\rho dt)^{(t_1-0)/dt}\right] \rho dt \\
& \left\{\delta_{t_1+dt,t_2} \xi + \right. \\
& \quad \left. (1-\delta_{t_1+dt,t_2}) (1-\xi) \left[(1-\rho dt)^{(t_2-t_1)/dt-2}\right] \rho dt\right\} \cdots \\
& \left\{\delta_{t_{N-1}+dt,t_N} \xi + \right. \\
& \quad \left. \left(1-\delta_{t_{N-1}+dt,t_N}\right) (1-\xi) \left[(1-\rho dt)^{(t_N-t_{N-1})/dt-2}\right] \rho dt\right\} \\
& (1-\xi) \left[(1-\rho dt)^{(T-t_N)/dt-2}\right],
\end{aligned}
$$

where $\delta_{t_n+dt,t_{n+1}}$ is Kronecker's delta. In the limit that $dt \to 0$, this simplifies to the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = e^{-\rho T}\rho dt \left\{ \prod_{n=1}^{N-1} \left[\delta_{t_n+dt,t_{n+1}}\xi + (1-\delta_{t_n+dt,t_{n+1}})(1-\xi)\rho dt\right] \right\} (1-\xi) \tag{S8}$$

$$= e^{-\rho T}\xi^{M}\left[(1-\xi)\rho dt\right]^{N-M}, \tag{S9}$$

where M is the number of times that $t_{n+1} - t_n = dt$. By taking the derivative of $\log \mathcal{L}(\boldsymbol{\theta})$ with respect to each of the parameters and setting the results equal to zero, it is straightforward to see that the uncensored likelihood function for a cascading Poisson process is optimized when the best-estimate parameters are specified by $\widehat{\xi} = M/N$ and $\widehat{\rho} = (N-M)/T$.

As in the case of the homogeneous Poisson process, when the data are interval censored we must instead estimate the parameters from the censored likelihood, Eq. (S7), which depends on the probability $\Pr(N_{T_\star}|\boldsymbol{\theta})$ of observing $N_{T_\star}$ events during a time window of $T_\star = 1$ day. Assuming that our data are generated by a cascading Poisson process, $\Pr(N_{T_\star}|\boldsymbol{\theta})$ is obtained by marginalizing $\Pr\big(t_1, t_2, \ldots, t_{N_{T_\star}}|\boldsymbol{\theta}\big)$ over all possible configurations of an ordered set of events occurring during this interval. If there are no events ($N_{T_\star} = 0$), then in the limit that $dt \to 0$ we are trivially left with

$$\begin{aligned}
\Pr(N_{T_\star} = 0|\boldsymbol{\theta}) &= \prod_{k=0/dt}^{T_\star/dt-1} \Pr_{\circ\circ}(kdt) \\
&= (1-\rho dt)^{T_\star/dt} \\
&= e^{-\rho T_\star},
\end{aligned} \tag{S10}$$

and if there are some events ($N_{T_\star} > 0$), we have from Eq. (S8)

$$\begin{aligned}
\Pr(N_{T_\star}|\boldsymbol{\theta}) = &\int_0^{T_\star}\int_{t_1}^{T_\star}\cdots\int_{t_{N_{T_\star}-1}}^{T_\star} e^{-\rho T_\star}\rho dt_1 \left\{ \prod_{n=1}^{N_{T_\star}-1} \left[\delta_{t_n+dt,t_{n+1}}\xi + (1-\delta_{t_n+dt,t_{n+1}})(1-\xi)\rho dt_{n+1}\right] \right\} (1-\xi) \\
&= e^{-\rho T_\star}\int_0^{T_\star}\int_{t_1}^{T_\star}\cdots\int_{t_{N_{T_\star}-2}}^{T_\star} (1-\xi)\rho dt_1 \left\{ \prod_{n=1}^{N_{T_\star}-2} \left[\delta_{t_n+dt,t_{n+1}}\xi + (1-\delta_{t_n+dt,t_{n+1}})(1-\xi)\rho dt_{n+1}\right] \right\} \\
&\quad \int_{t_{N_{T_\star}-1}}^{T_\star} \left[\delta_{t_{N_{T_\star}-1}+dt,t_{N_{T_\star}}}\xi + (1-\delta_{t_{N_{T_\star}-1}+dt,t_{N_{T_\star}}})(1-\xi)\rho dt_{N_{T_\star}}\right]
\end{aligned}$$

$$\Pr(N_{T_\star}|\boldsymbol{\theta}) = e^{-\rho T_\star} \int\limits_0^{T_\star} \int\limits_{t_1}^{T_\star} \cdots \int\limits_{t_{N_{T_\star}-2}}^{T_\star} (1-\xi)\rho dt_1 \left\{ \prod_{n=1}^{N_{T_\star}-2} \left[\delta_{t_n+dt,t_{n+1}} \xi + (1-\delta_{t_n+dt,t_{n+1}})(1-\xi)\rho dt_{n+1} \right] \right\}$$
$$\left[\xi + \int\limits_{t_{N_{T_\star}-1}}^{T_\star} (1-\xi)\rho dt_{N_{T_\star}} \right]$$
$$= e^{-\rho T_\star} \int\limits_0^{T_\star} \int\limits_{t_1}^{T_\star} \cdots \int\limits_{t_{N_{T_\star}-2}}^{T_\star} (1-\xi)\rho dt_1 \left\{ \prod_{n=1}^{N_{T_\star}-2} \left[\delta_{t_n+dt,t_{n+1}} \xi + (1-\delta_{t_n+dt,t_{n+1}})(1-\xi)\rho dt_{n+1} \right] \right\}$$
$$\left[\xi + (1-\xi)\rho(T_\star - t_{N_{T_\star}-1}) \right]$$
$$\vdots$$
$$= e^{-\rho T_\star}(1-\xi)\rho T_\star \sum_{n=0}^{N_{T_\star}-1} \binom{N_{T_\star}-1}{n} \frac{\xi^n \left[(1-\xi)\rho T_\star\right]^{N_{T_\star}-1-n}}{(n+1)!}. \tag{S11}$$

Taking Eqs. (S10–S11) together, we see that

$$\Pr(N_{T_\star}|\boldsymbol{\theta}) = \begin{cases} e^{-\rho T_\star} & N_{T_\star} = 0 \\ e^{-\rho T_\star} Q(N_{T_\star}-1; \boldsymbol{\theta}, T_\star) & N_{T_\star} > 0 \end{cases} \tag{S12}$$

where the polynomial

$$Q(N; \boldsymbol{\theta}, T_\star) \;\; = \;\; (1-\xi)\rho T_\star \sum_{n=0}^{N} \binom{N}{n} \frac{\xi^n \left[(1-\xi)\rho T_\star\right]^{N-n}}{(n+1)!} \tag{S13}$$

accounts for the various ways that the $N = N_{T_\star} - 1$ events during the time segment of duration $T_\star$ time units can be grouped into cascades of activity. Estimating the parameters of the cascading Poisson process from the censored likelihood function is analytically intractable. Instead, we estimate the parameters of the cascading Poisson process by numerically maximizing the corresponding censored likelihood function, Eq. (S7), for the cascading Poisson process.

S5 Monte Carlo hypothesis testing

Given a model $\mathcal{M}$ with parameters $\boldsymbol{\theta}_i$, we use Monte Carlo hypothesis testing to determine whether the model can be rejected during each time segment $[T_i, T_{i+1})$ of duration $\Delta T_i = T_{i+1} - T_i$ (*S16, S15*). The Monte Carlo hypothesis testing procedure is as follows. First, we calculate the best-estimate parameters $\widehat{\boldsymbol{\theta}}_i$ for model $\mathcal{M}$ using maximum likelihood estimation. Second, we compute the test statistic $\mathcal{S}$ (detailed below) between the model $\mathcal{M}(\widehat{\boldsymbol{\theta}}_i)$ and the empirical data $\mathcal{D}_i$ during that time segment $[T_i, T_{i+1})$. We next generate a synthetic data set $\mathcal{D}_s$ from model $\mathcal{M}(\widehat{\boldsymbol{\theta}}_i)$ over the same time segment $[T_i, T_{i+1})$ using the best-estimate parameters $\widehat{\boldsymbol{\theta}}_i$, and we treat the synthetic data exactly the same as we treated the empirical data: first, we calculate the best-estimate parameters $\widehat{\boldsymbol{\theta}}_s$ for model $\mathcal{M}$ from maximum likelihood estimation; second, we compute the test statistic $\mathcal{S}_s$ between the model $\mathcal{M}(\widehat{\boldsymbol{\theta}}_s)$ and the synthetic data $\mathcal{D}_s$. We generate synthetic data sets $\mathcal{D}_s$ and their corresponding synthetic test statistics $\mathcal{S}_s$ until we accumulate an ensemble of 10,000 Monte Carlo test statistics $\{\mathcal{S}_s\}$. Finally, we calculate a two-tailed p-value with a precision of 10^{-4} by computing $\Pr(|\mathcal{S}_s - \langle\mathcal{S}_s\rangle| > |\mathcal{S} - \langle\mathcal{S}_s\rangle|)$ where $\langle\mathcal{S}_s\rangle$ is a suitably chosen centroid of the distribution of synthetic test-statistics. As is customary in hypothesis testing, we reject the model $\mathcal{M}$ during time segment $[T_i, T_{i+1})$ if the p-value is less than a threshold value. We select a p-value threshold of 0.05; that is, if less than 5% of the synthetic data sets exhibit deviations in the test statistic that are larger than those observed empirically, the model is rejected for that time segment $[T_i, T_{i+1})$.

Testing a model over a particular time segment $[T_i, T_{i+1})$ introduces two challenges to hypothesis testing. First, an important consideration in Monte Carlo hypothesis testing is that we must use a distribution for which both the empirical and synthetic data sets have the same number of observations. Since our synthetic data is generated during a specified time segment $[T_i, T_{i+1})$, we can not use the inter-event time distribution because each synthetic time series is not guaranteed to have the same number of events as the empirical time series. Instead, we assess the consistency of our model with the empirical data by comparing the distribution $\Pr(N_{T_\star}|\boldsymbol{\theta})$ of the number of events $N_{T_\star}$ during a time period of a specified duration $T_\star$. We choose a duration of $T_\star = 1$ week as this seems to be a reasonable time scale for human activity (*S15*), so both the synthetic and empirical distributions $\mathrm{Pr}(N_{T_\star}|\boldsymbol{\theta})$ have $\Delta T_i/T_\star$ observations. We have confirmed that our results are insensitive to the specific choice of $T_\star$ provided that $T_\star \ll \Delta T_i$.

Second, since we use the distribution $\mathrm{Pr}(N_{T_\star}|\boldsymbol{\theta})$ of the number of events $N_{T_\star}$ during a time period of a duration $T_\star = 1$ week—a discrete distribution—it is important to use a test statistic $\mathcal{S}$ that is appropriate for testing discrete distributions. We use the χ^2 test statistic. An important consideration in using the χ^2 test statistic is that one must bin the observations and expected observations according to model $\mathcal{M}(\widehat{\boldsymbol{\theta}}_i)$ in a meaningful way. We bin $\mathrm{Pr}(N_{T_\star}|\boldsymbol{\theta})$ such that each bin has at least one expected observation according to model $\mathcal{M}(\widehat{\boldsymbol{\theta}}_i)$, which prevents observations that are exceptionally rare from dominating our statistical test and skewing our results.

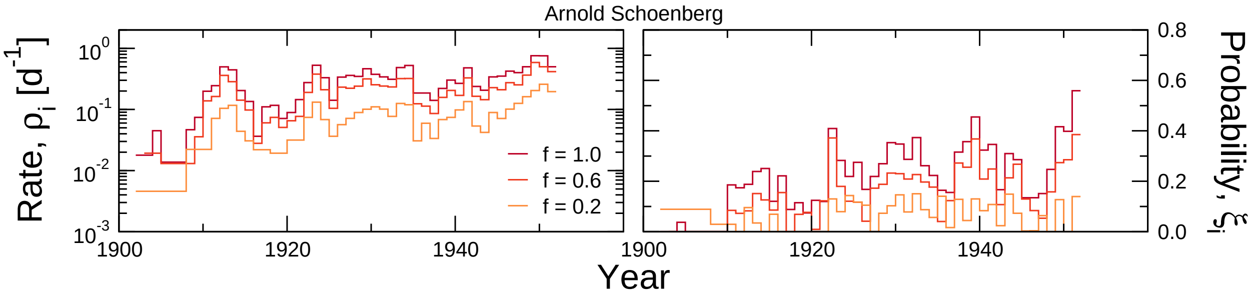


Fig. S1: Cascading Poisson process best-estimate parameters $\boldsymbol{\theta} = \{\rho_i, \xi_i\}$ during each time segment for Arnold Schoenberg when only a fraction of the original letters are returned to the centralized database. We include here the parameter estimates for the when 100%, 60%, and 20% of all letters are returned to the centralized database.

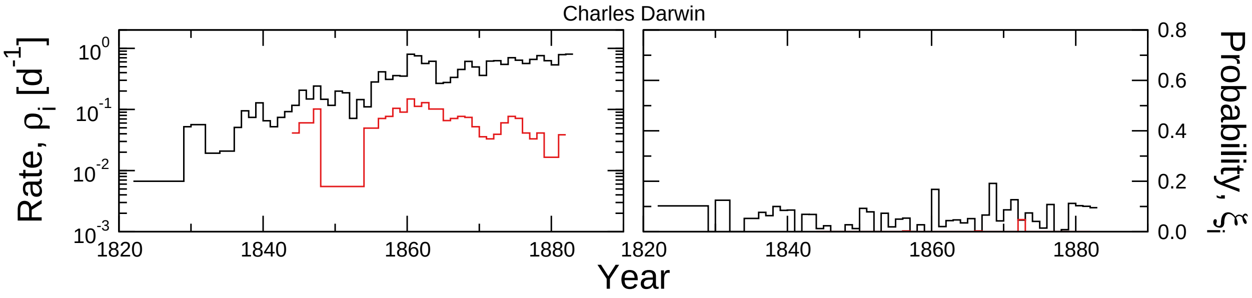


Fig. S2: Cascading Poisson process best-estimate parameters $\boldsymbol{\theta} = \{\rho_i, \xi_i\}$ during each time segment for Charles Darwin when we consider all of his correspondence (black line) or only his correspondence to J.D. Hooker (red line).

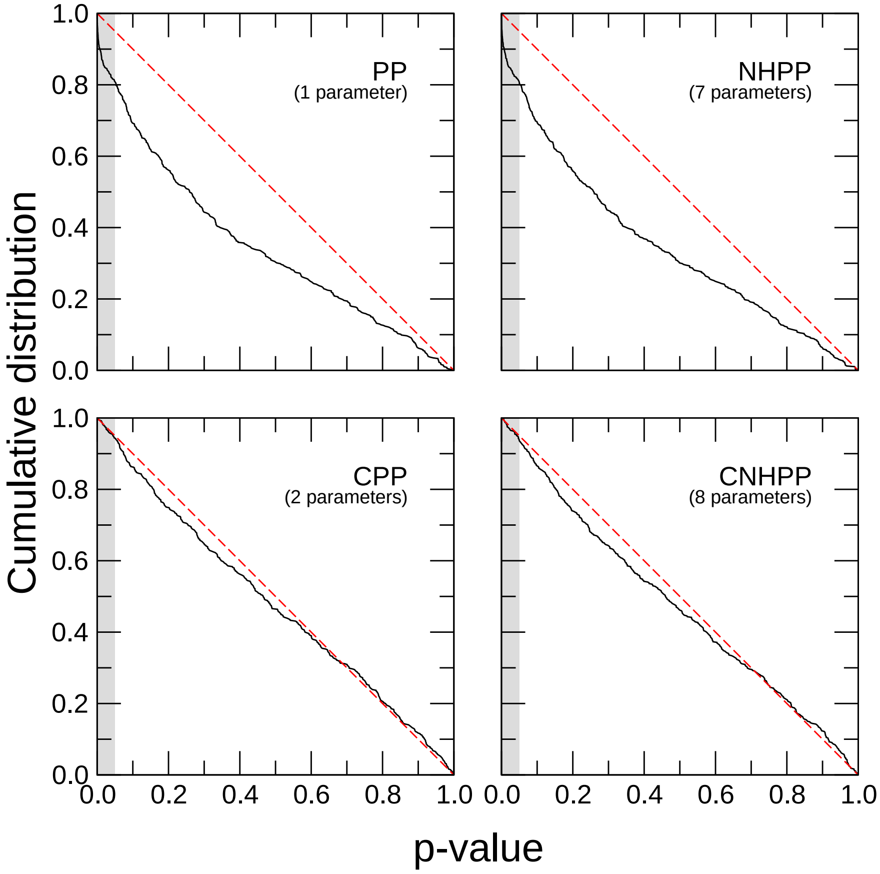
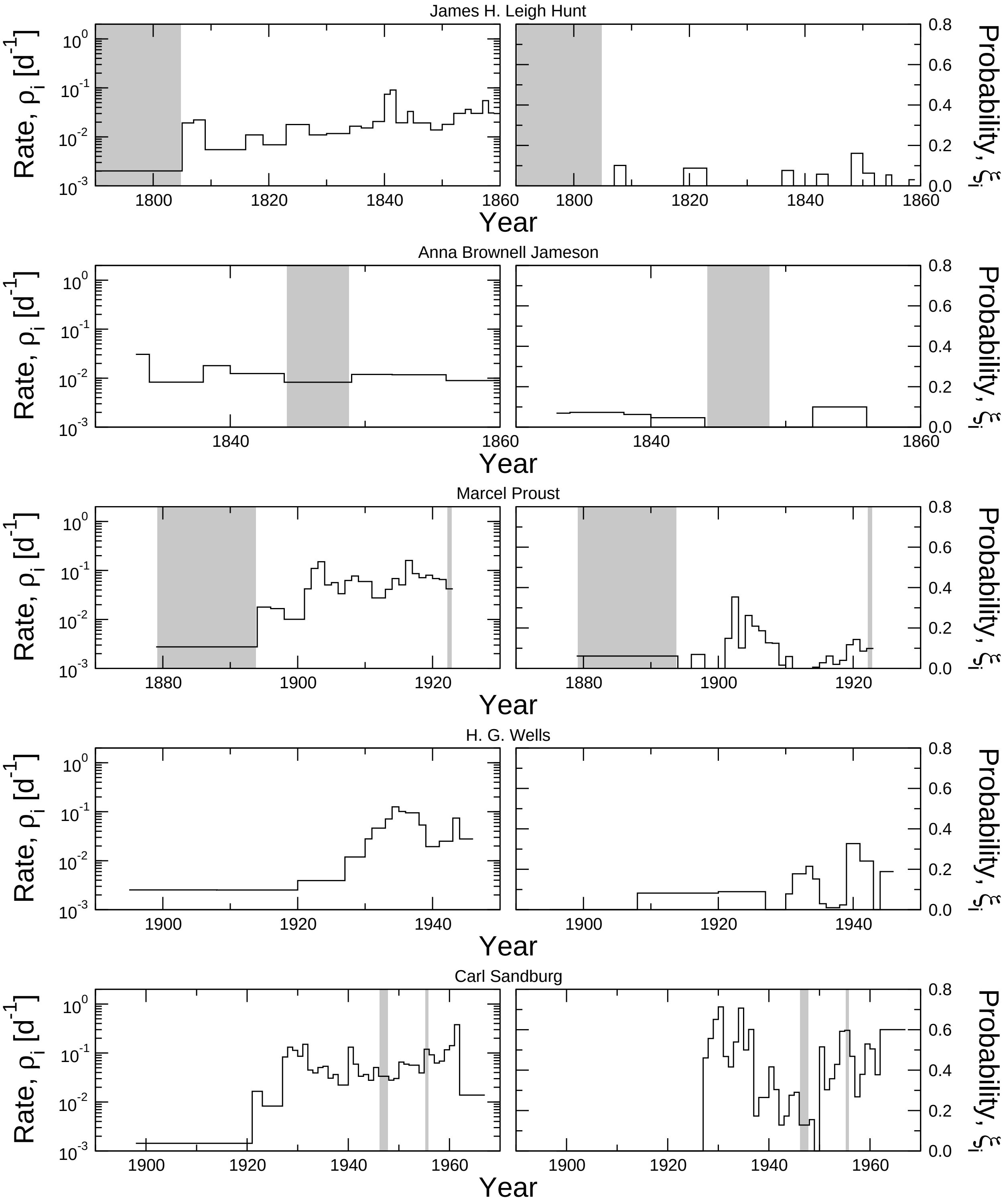
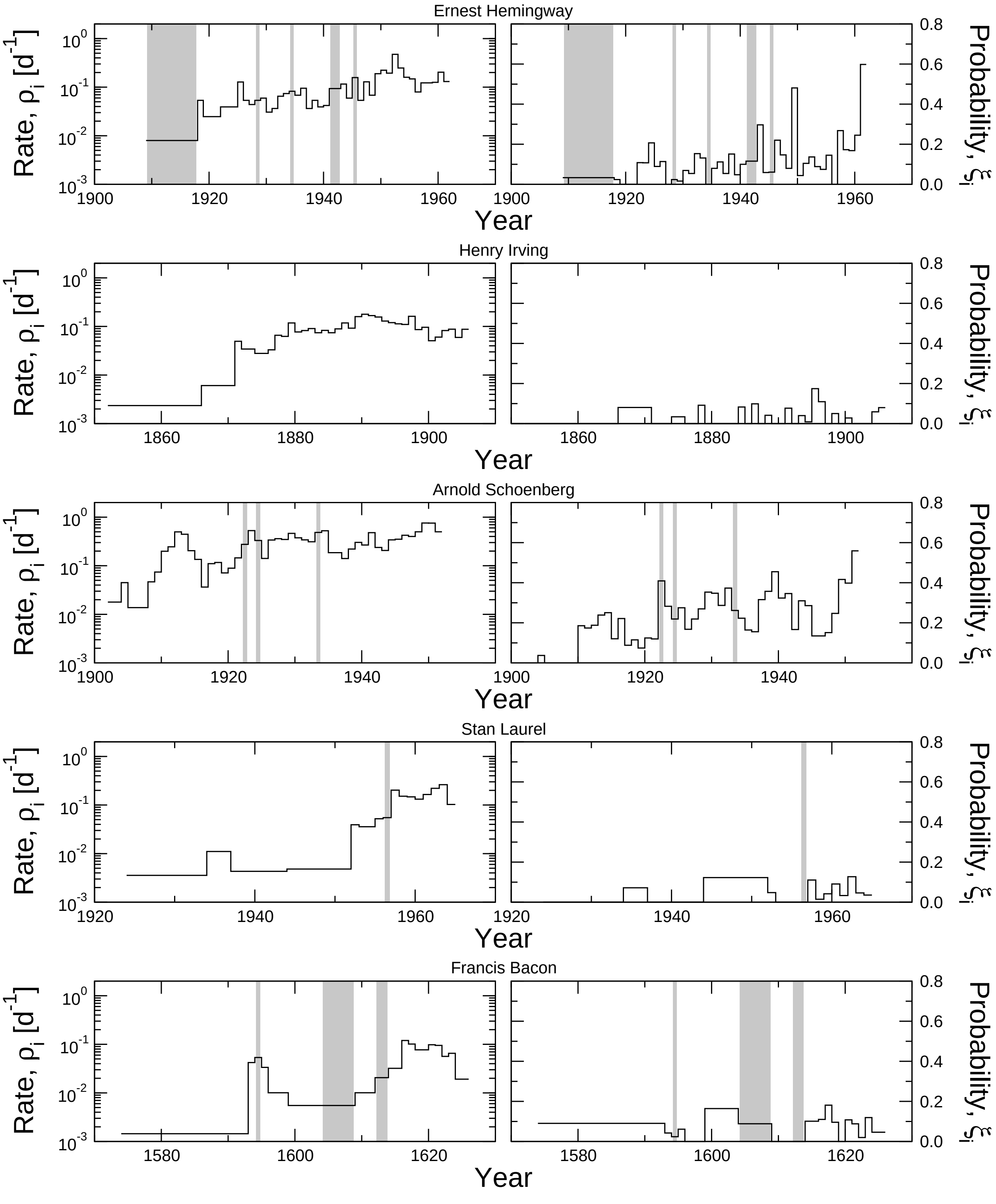
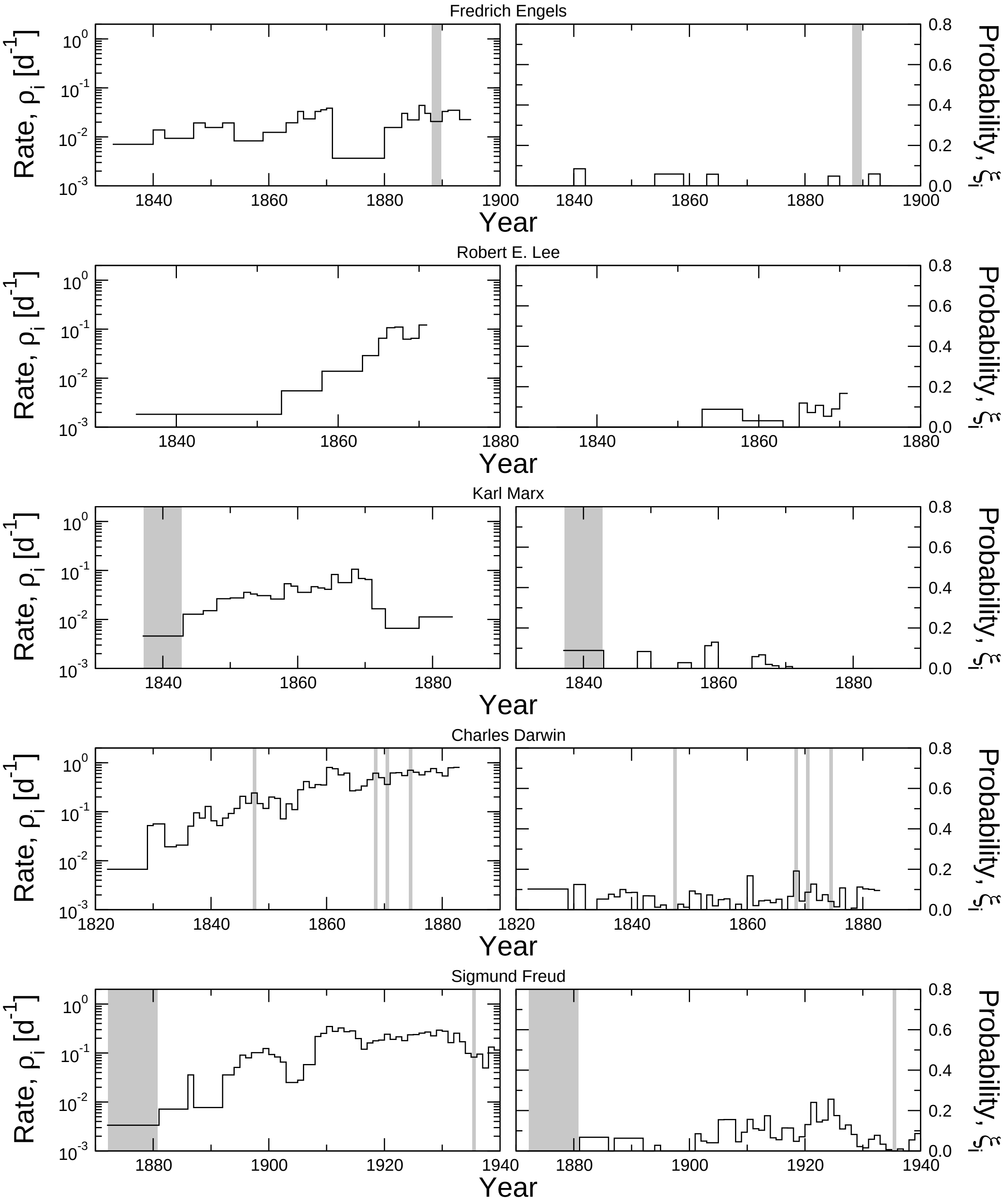


Fig. S3: Cumulative distribution of p-values from the Monte Carlo hypothesis tests for all 16 letter correspondents during each time segment $[T_i, T_{i+1})$ (black line) for each model under consideration: a homogeneous Poisson process (PP), a non-homogeneous Poisson process (NHPP), a cascading Poisson process (CPP), and a cascading non-homogeneous Poisson process (CNHPP). We reject a model during a particular time segment $[T_i, T_{i+1})$ if the p-value is less than 0.05 (grey shaded region). Note that if the data were drawn from one of these models, we would expect a uniform distribution of p-values (dashed red line). Since this is very nearly the case for the cascading Poisson process and the cascading non-homogeneous Poisson process, this provides additional evidence that these models are consistent with letter correspondence patterns.


James H. Leigh Hunt
Anna Brownell Jameson
Marcel Proust
H. G. Wells
Carl Sandburg
Rate, ρ_i [d^{-1}]
Probability, ξ_i
Year
10^{0}
10^{-1}
10^{-2}
10^{-3}
0.0
0.2
0.4
0.6
0.8
1800
1820
1840
1860
1880
1900
1920
1940
1960


Ernest Hemingway
Henry Irving
Arnold Schoenberg
Stan Laurel
Francis Bacon
Rate, ρ_i [d^{-1}]
Probability, ξ_i
Year
10^{0}
10^{-1}
10^{-2}
10^{-3}
0.8
0.6
0.4
0.2
0.0
1900
1920
1940
1960
1860
1880
1580
1600
1620


Fredrich Engels
Robert E. Lee
Karl Marx
Charles Darwin
Sigmund Freud
Rate, ρ_i [d^{-1}]
Probability, ξ_i
Year
10^0
10^{-1}
10^{-2}
10^{-3}
0.8
0.6
0.4
0.2
0.0
1820
1840
1860
1880
1900
1920
1940

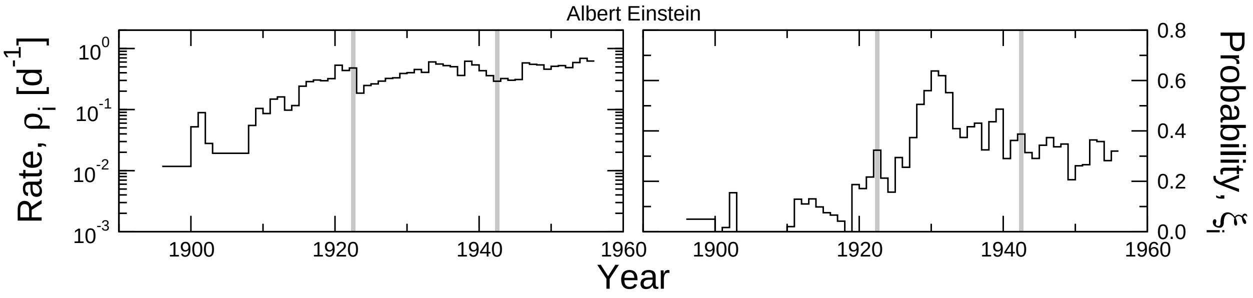
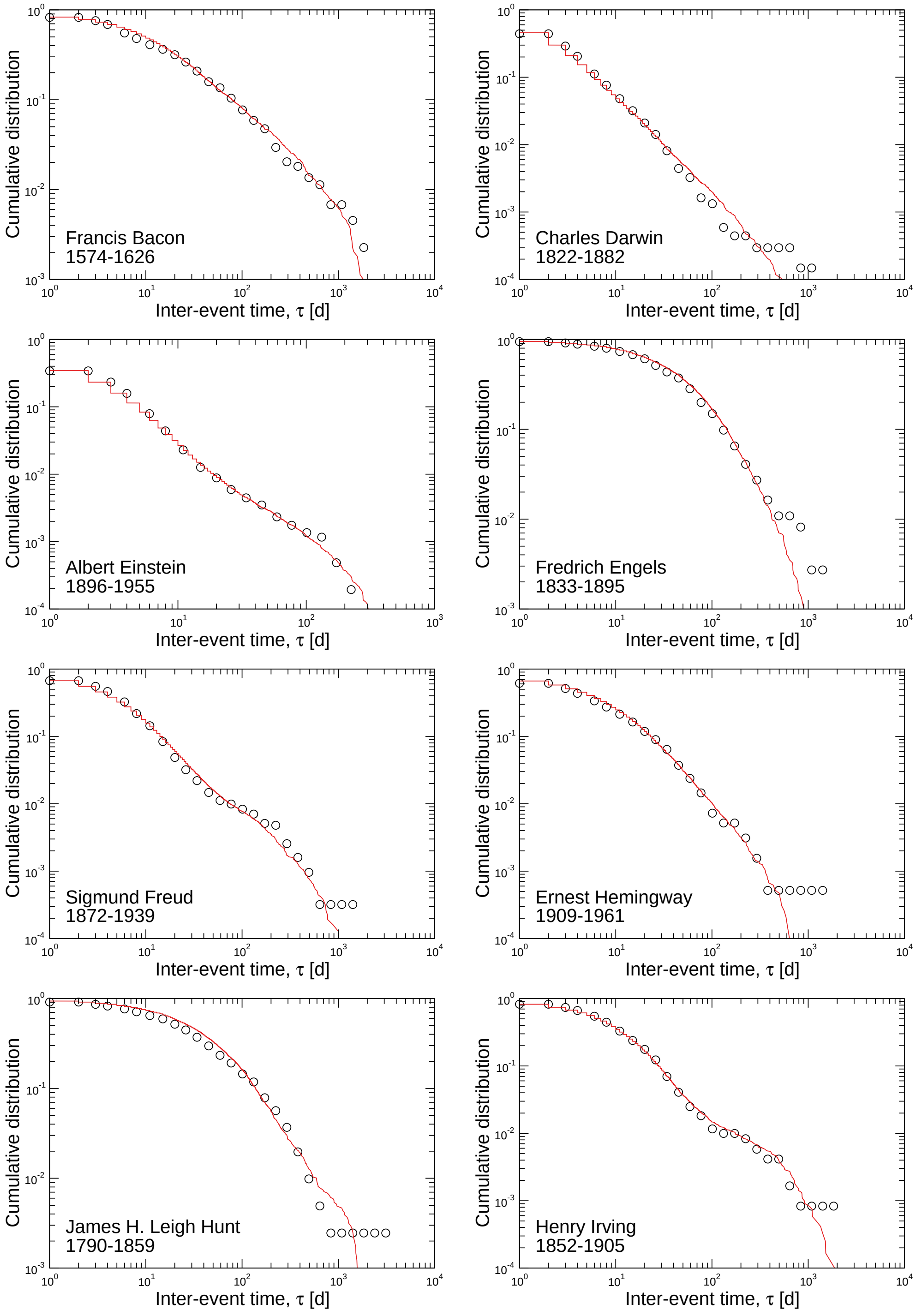


Fig. S4: Parameter estimates for a cascading Poisson process for all 16 writers, performers, politicians, and scientists under consideration. We estimate the parameters $\boldsymbol{\theta}_i = \{\rho_i, \xi_i\}$ during each time segment $[T_i, T_{i+1})$ for a cascading Poisson process by maximum likelihood. Grey shaded regions denote time segments during which a cascading Poisson process is rejected by Monte Carlo hypothesis testing.


Francis Bacon
1574-1626
Charles Darwin
1822-1882
Albert Einstein
1896-1955
Fredrich Engels
1833-1895
Sigmund Freud
1872-1939
Ernest Hemingway
1909-1961
James H. Leigh Hunt
1790-1859
Henry Irving
1852-1905
Cumulative distribution
Inter-event time, τ [d]
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{1}
10^{2}
10^{3}
10^{4}

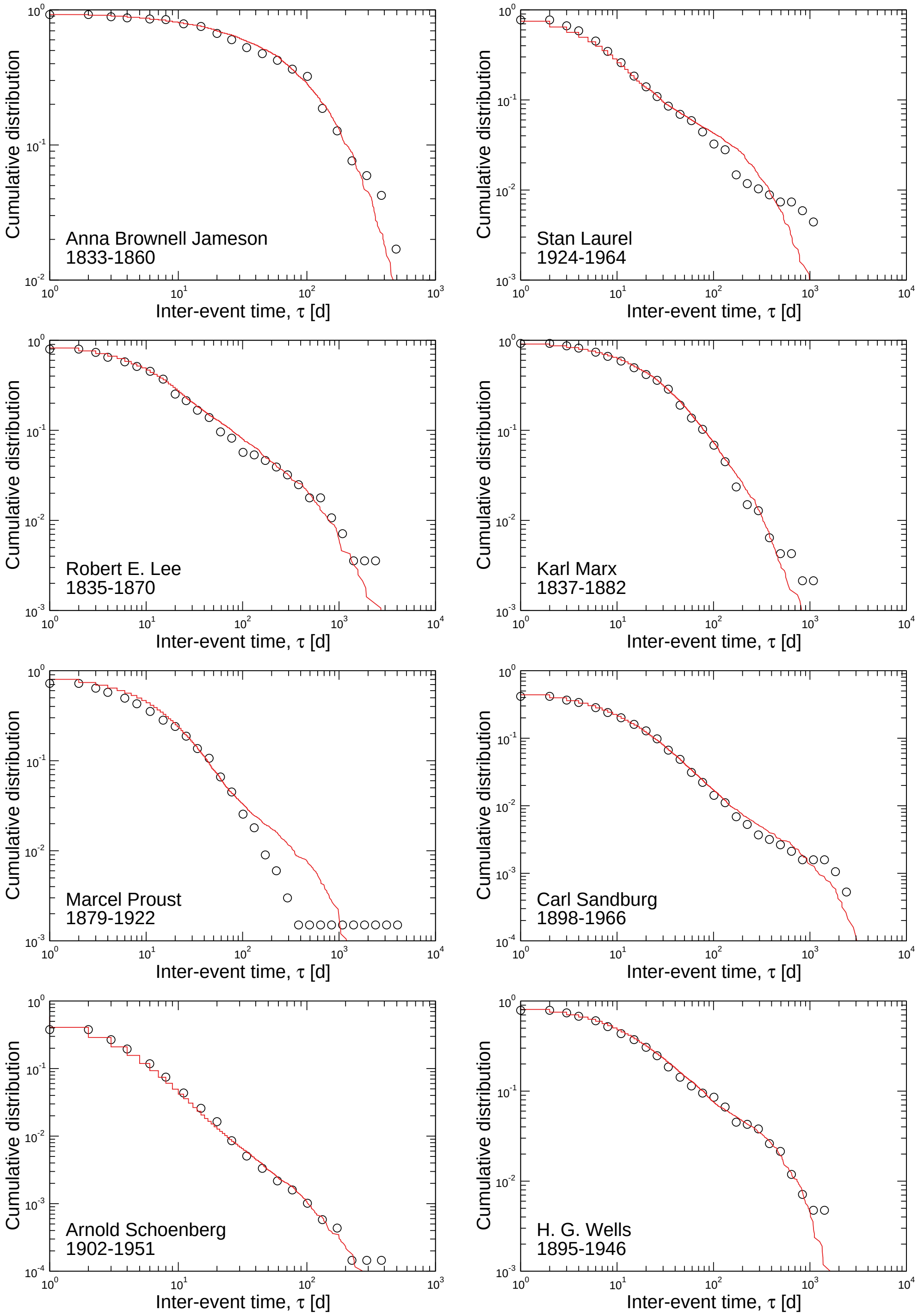


Fig. S5: Comparison of the inter-event time distribution for all 16 individuals (circles) and the predictions of a non-stationary cascading Poisson process (red line). The predictions of the non-stationary cascading Poisson process are estimated numerically.

	Number of sent letters			
Individual (Reference)	Before processing	After processing	Number of recipients	Letters per year
Francis Bacon (*S1*)	673	443	174	8.36
James H. Leigh Hunt (*S2*)	604	408	219	5.83
Charles Darwin (*S3*)	$7,595$	$6,785$	661	111
Anna Brownell Jameson (*S4*)	302	119	58	4.25
Friedrich Engels (*S5*)	413	369	70	5.86
Robert E. Lee (*S6*)	285	282	213	7.83
Karl Marx (*S5*)	491	469	72	10.2
Henry Irving (*S7*)	$1,621$	$1,205$	15	22.3
Sigmund Freud (*S8*)	$3,162$	$3,130$	168	46.0
Marcel Proust (*S9*)	670	668	135	15.2
H. G. Wells (*S9*)	$1,088$	422	$1,041$	8.12
Albert Einstein (*S10*)	$14,512$	$10,319$	$5,207$	172
Carl Sandburg (*S11*)	$2,971$	$1,894$	$2,771$	27.4
Arnold Schoenberg (*S9*)	$7,925$	$6,899$	$1,848$	138
Ernest Hemingway (*S12*)	$2,363$	$1,934$	532	36.5
Stan Laurel (*S13*)	693	685	157	16.7

Tbl. S1: Summary of the letter correspondence records for the 16 individuals under consideration. For each individual, we note the total number of sent letters before and after processing, the number of recipients and the average number of letters per year.

Fraction of letters	Number of segments	95% CI	Number of rejections
1.0	47	$[0, 5]$	3
0.9	47	$[0, 5]$	3
0.8	47	$[0, 5]$	3
0.7	45	$[0, 5]$	1
0.6	45	$[0, 5]$	2
0.5	44	$[0, 5]$	5
0.4	44	$[0, 5]$	1
0.3	42	$[0, 5]$	0
0.2	40	$[0, 4]$	1
0.1	32	$[0, 4]$	1

Tbl. S2: Summary of the hypothesis testing results for Arnold Schoenberg when only a fraction of the in the centralized data base are considered. As the fraction of letters considered decreases, more time segments must be merged such that at least 10 events occur within each time segment.

Individual	Number of segments	95% CI	Number of rejections PP	NHPP	CPP	CNHPP
Francis Bacon	19	$[0,3]$	**4**	**4**	3	3
James H. Leigh Hunt	25	$[0,3]$	2	3	1	1
Charles Darwin	52	$[0,5]$	**7**	**7**	4	4
Anna Brownell Jameson	8	$[0,2]$	1	1	1	1
Friedrich Engels	24	$[0,3]$	1	2	1	2
Robert E. Lee	10	$[0,2]$	1	1	0	1
Karl Marx	25	$[0,3]$	1	1	1	0
Henry Irving	35	$[0,4]$	1	1	0	1
Sigmund Freud	49	$[0,5]$	2	2	2	3
Marcel Proust	25	$[0,3]$	2	2	2	1
H. G. Wells	16	$[0,2]$	**3**	1	0	0
Albert Einstein	54	$[0,6]$	**21**	**22**	2	4
Carl Sandburg	37	$[0,4]$	**15**	**15**	2	2
Arnold Schoenberg	47	$[0,5]$	**23**	**23**	3	3
Ernest Hemingway	42	$[0,5]$	**7**	**7**	5	4
Stan Laurel	17	$[0,3]$	2	2	1	1

Tbl. S3: Summary of the letter correspondence records and hypothesis testing results for the 16 individuals. For each individual, we note the number of time segments $[T_i, T_{i+1})$ with at least 10 letters per time segment, the 95% confidence interval (CI) bounds on a binomial model with $p = 0.05$, and the number of rejections based on our Monte Carlo hypothesis testing procedure for each of the models we test: a homogeneous Poisson process (PP), a non-homogeneous Poisson process (NHPP), a cascading Poisson process (CPP), and a cascading non-homogeneous Poisson process (CNHPP). The number of rejections is highlighted in bold if the model is not consistent with the data.

References and Notes

S1. Francis Bacon Correspondence Project, http://webapps.qmul.ac.uk/cell/Bacon/chronological_index.html.

S2. Iowa Digital Library: Leigh Hunt Digital Collection, http://tinyurl.com/ad4mj8.

S3. The Darwin Correspondance Project, http://www.darwinproject.ac.uk.

S4. The Victorian Women Writers' Letters Project, http://www.lib.sfu.ca/cgi-bin/edocs/SearchVWWLP.

S5. Marx & Engels Internet Archive, http://www.marxists.org/archive/marx.

S6. Robert E. Lee Collection, http://home.wlu.edu/ stanleyv/pentrans.htm.

S7. Henri Irving Correspondence, http://www.henryirving.co.uk.

S8. Research Centre, http://www.freud.org.uk.

S9. The Rare Book & Manuscript Library, University of Illinois, http://www.library.uiuc.edu/rbx/manuscript.htm.

S10. The Einstein Papers Project, http://www.einstein.caltech.edu.

S11. The Rare Book & Manuscript Library, http://www.library.uiuc.edu/rbx/SandburgConnemara.html.

S12. Hemingway Archives – John F. Kennedy Presidential Library & Museum, http://www.jfklibrary.org/Historical Resources/Hemingway+Archive.

S13. The Stan Laurel Correspondence Archive Project, http://www.lettersfromstan.com.

S14. B. Chapman, J. Chang, *ACM SIGBIO Newslett.* **20**, 15 (2000).

S15. R. D. Malmgren, D. B. Stouffer, A. E. Motter, L. A. N. Amaral, *Proc. Natl. Acad. Sci. USA* **105**, 18135 (2008).

S16. R. B. D'Agostino, M. A. Stephens, *Goodness-of-Fit Techniques* (Marcel Kekker, Inc., 1986).