## Supporting Information

## Mathematical Properties of the DDa+PA model of PIN evolution.

We introduce a more rigorous terminology for the mixture models to PIN evolution presented in Materials and Methods, and show in particular that (1) DDa may explain any protein network topology and (2) computing the likelihood of a PIN generated by DDa is computationally infeasible. Here we focus on the evolution process only and do not consider the sampling component of our model that account for incompleteness in observed PINs.

Let $G=(V, E)$ denote a graph with node set $V$ and edge set $E$, where $e \in E$ if there are nodes $u, v \in V$ such that $e$ is a link between $u$ and $v$. We always take $G$ to be undirected without multiple edges, self-loops and labelled.

Definition S. 1 A randomly growing graph is a Markov chain $\left\{G_{t}\right\}_{t=t_{0}}^{\infty}$ such that $G_{t}$ is a subset of the nodes of $G_{t+1}$. Denote the transition probability from $G_{t}$ to $G_{t+1}$ by $P\left(G_{t+1} \mid G_{t}\right)$ and consider the transition kernel $P:=P(\cdot \mid \cdot)$. A randomly growing graph together with a stopping rule $T$ defines an evolution graph $\left\{G_{t}\right\}_{t=t_{0}}^{T}$. Equivalently, a transition kernel, an initial graph and a stopping rule uniquely define an evolution graph.

Note that the above definition readily implies that only one node be added per time. Since graphs are labelled, the last node added in $G_{t}$ is uniquely determined. Here we consider deterministic stopping rules only. In this text we focus on the transition kernel $P$ that corresponds to evolution dynamics. It is straighforward to determine the transition probabilities for the mixture model presented in the main text; recall in particular that the node $u$ for duplication is chosen uniformly. Denote the growth parameters of this mechanism by $\theta=\left(\delta_{\text {Div }}, \delta_{\text {Att }}, \alpha\right)$.

Lemma S. 2 Given $G_{t}$ and $G_{t+1}$, assume that node $u$ is the parent of the new node $v$ in $G_{t+1}$. We have that

$$
\begin{equation*}
P\left(G_{t+1} \mid G_{t}, u, \theta\right)=\alpha P A(u, v)+(1-\alpha) D D a\left(u, v, \delta_{D i v}, \delta_{A t t}\right) \tag{S.10}
\end{equation*}
$$

where $P A(u, v):=\mathbf{1}(e(u, v)=1 \wedge d(v)=1)$,

$$
D D a\left(u, v, \delta_{D i v}, \delta_{A t t}\right):=C\left[\delta_{A t t}^{e}(u, v)\left(1-\delta_{A t t}\right)^{1-e(u, v)}\right]\left(\left[1-\delta_{D i v}\right]^{2 N_{11}(u, v)}\left[\delta_{D i v}\left(1-\delta_{D i v}\right)\right]^{N_{10}(u, v)}\right),
$$

$e(u, v)$ is 1 if there is an edge between $u$ and $v$ and 0 otherwise, $d(v)$ denotes the degree of node $v, N_{11}(u, v):=$ $\# N(u) \cap N(v)$ and $N_{10}(u, v):=\# N(u) \cup N(v)-\# N(u) \cap N(v)-e(u, v)$. The constant $C$ of normalization is $\left(1-\delta_{D i v}{ }^{2}\right)^{N_{11}(u v)+N_{10}(u, v)}$.

Proof: Suppose that $x$ in $G_{t}$ duplicated and diverged to the nodes $u$ and $v$ in $G_{t+1}$. Let $N(x)$ be the set of nodes in $G_{t}$ that are adjacent to $x$, and denote for all $u, v$ the set numbers $M_{11}(u, v):=\#\{y \in N(x) \mid y \in N(u) \wedge y \in N(v)\}$, $M_{01}(u, v):=\#\{y \in N(x) \mid y \notin N(u) \wedge y \in N(v)\}$ and $M_{10}(u v):=\#\{y \in N(x) \mid y \in N(u) \wedge y \notin N(v)\}$. Since edges duplicate and diverge uniformly and independently, the transition probability of duplication-divergence discarding links between the parent and child is proportional to

$$
\left[\left(1-\delta_{\text {Div }}\right)\left(1-\delta_{\text {Div }}\right)\right]^{M_{11}(u, v)}\left[\delta_{\text {Div }}\left(1-\delta_{\text {Div }}\right)\right]^{M_{01}(u, v)}\left[\left(1-\delta_{\text {Div }}\right) \delta_{\text {Div }}\right]^{M_{10}(u, v)}
$$

If we denote $N_{11}(u, v)$ and $N_{10}(u, v)$ as above, then the above expression simplifies to

$$
\left.\left[1-\delta_{\text {Div }}\right)\right]^{2 N_{11}(u, v)}\left[\delta_{\text {Div }}\left(1-\delta_{\text {Div }}\right)\right]^{N_{10}(u, v)}
$$

and (S.10) follows easily.

Corollary S. 3 The transition probability from $G_{t}$ to $G_{t+1}$ given the set $\theta$ of growth parameters is

$$
\begin{equation*}
P\left(G_{t+1} \mid G_{t}, \theta\right)=\frac{1}{t} \sum_{u} P\left(G_{t+1} \mid G_{t}, u, \theta\right) \tag{S.11}
\end{equation*}
$$

If not stated otherwise, the (standard) mixture model in this article is the evolution graph with transition probabilities (S.11), initial graph $G_{0}$ that connects two nodes with one link and stopping time $T \ll \infty$. For H. pylori and P. falciparum, we take $T=1500$ and $T=5300$ respectively.

Definition S. 4 For a graph $G_{t}$ and a node $v$ in $G_{t}$, denote $G_{t}$ where node $v$ is removed by $G_{t}(-v)$. $v$ is said to be removable if $P\left(G_{t} \mid G_{t}(-v), \theta\right)>0$ for some $\theta$. If $G_{t}$ contains removable nodes, then it is called reducible, otherwise irreducible. An evolution graph is reducible, if $P\left(G_{t} \mid G_{1}\right)>0$ for some $\theta$ and any graph $G_{t}$, where $G_{1}$ is the graph consisting of one node.

Lemma S. 5 Consider the transition kernel $P$ of the mixture model. For any nontrivial graph $G_{t}$ and a node $v$ of $G_{t}$

$$
P\left(G_{t} \mid G_{t}(-v)\right)>0
$$

where we set $\delta_{A t t}$ and $\delta_{\text {Div }}$ different from 0,1 and $\alpha<0$.
Proof: $\quad$ Since $P$ does not produce a graph with nodes of degree 0 , we may choose a node $w \neq v$ in $G_{t}$. If $G_{t}$ contains only $v$ and $w$, then there must be an edge between them and clearly $p\left(G_{2} \mid G_{1}, v, \theta\right)=1-\delta_{\text {Att }}>0$. Let us now assume that $G_{t}$ contains more than two nodes. Consider $N:=N(v) \cup N(w) \backslash\{v, w\}$. Then, we may construct $G_{t}(-v)$ as the graph that results after replacing $v, w$ and $N(v), N(w)$ with its common ancestor $u, N$.

Applying Lemma S. 5 repeatedly, gives the following Corollary.
Corollary S. 6 The mixture model is reducible.
Corollary S. 6 guarantees that the the mixture model can explain any graph $G$ topologically. In particular, Lemma S. 5 is based on properties of DDa , and hence a fortiori DDa can explain any graph topologically. It also shows that every node is removable and hence there are $T$ ! ways to reduce any graph $G_{t}$ with DDa . This is significantly less for the transition kernel presented in [19], making the inference scheme therein computationally feasible.

## Summarizing aspects of PINs



Figure S.1: The PIN data sets generated by process (1) were dominated by stochastic effects. 1000 networks to $H$. pylori (grown to 1500 nodes and subsampled to 675 ) are generated with the parameter $\theta=(0.32,0.02,0.15)$, and the squared errors between each summary and the mean summary are recorded. The frequency of cases such that the squared error is greater than values on the abscissa is plotted for WR, ND, PL, ND and TRIA. In $20 \%$ of all cases, the squared error in TRIA is greater than 1000, while in all cases the squared error in ND is not larger than 1. Except for ND, large deviations are likely for all summaries, reflecting that stochastic effects dominate network summaries.

(a)

(b)

(c)

Figure S.2: Different parameters of model (1) leave distinguishable imprints on simulated PINs. We compared WR and NDfor $\alpha=0,0.2,1$ to the observed summaries of H. pylori (grey) by simulating 50 networks to $H$. pylori (grown to the number of open-reading frames, 1500, and subsampled to the observed network order, 675) with $\theta=(0.24,0.04, \alpha)$ for varying $\alpha$. For each withindistance $d$ and each node degree $k$ the interquantile range of $p(w r \leq d)$ and $p(k)$ for the 50 generated networks was drawn. (a) The interquantile ranges of WR for PINs generated by different parameters were clearly distinct, and the mixture model with $\alpha=0.2$ visually improved fit relative to DDa and PA. (b) On the same scale, the interquantile ranges of ND largely overlapped, indicating that ND might have significantly less power than WR to distinguish between different parameters. (c) On the log scale for $p(k)$, the interquantile ranges of ND generated by different parameters were again distinguishable, suggesting that the use of different distance metrics might play an important role in inference on protein network data.


Figure S.3: Mean summaries over larger ensembles of $B$ simulated PIN data sets have reduced variance, as exemplified here with DIAM. We computed the mean summary (red points) from $B=200,50,5$ networks to $H$. pylori (grown to 1500 nodes with $\theta=(0.28,0.03,0.21)$ and subsampled to 675 nodes). In each computation, the 50 networks were randomly chosen from the 200 networks and then the 5 networks randomly from the 50 networks. This procedure was repeated 100 times, and we report the density of the distance of the mean simulated DIAM to the observed DIAM for $B=200,50,5$. The average of these errors (vertical red line) and the range of one standard deviation (blue) are added. Clearly, the variance of the mean DIAM shrinks with increasing $B$, and similarly for all other summaries (not shown) with $\sqrt{B}$ according to the Central Limit Theorem (not shown).


Figure S.4: To compare the variability of the mean posterior summaries of $H$. pylori, we studied the density of standardized variation $\operatorname{cv}(\theta)$, described in Materials and Methods, on the grid $\theta \in[0.1,0.7] \times[0,0.5] \times[0.1,0.6]$ in steps of 0.025 . Computations were based on summaries taken from 1000 simulated PINs to H. pylori (grown to 1500 nodes and subsampled to 675 ). We plot the marginal $\operatorname{cv}(\alpha)$ against $\alpha$ for (a) summary statistics and (b) summary distributions. cv complements the information given by smd in Figure 1 to characterize the sensitivity and variability of the summary statistics. TRIA, FRAG and CC are extremely variable, offsetting their high standardized mean derivatives. ND is almost invariant to random fluctuations and to different parameters. Results for the other two parameters are very similar (not shown).


Figure S.5: Many proteins are missing in PINs, and, ideally, summaries should be robust under increasing incompleteness. For a given parameter $\theta$, we compared the curvature of decay of ND with those of PL and TRIA under decreasing $\rho$. The interquantile range and mean summaries for PL, $\overline{\mathrm{ND}}$ and TRIA are displayed (in blue) as a function of decreasing sampling fraction $\rho$. Computations were based on 250 networks to $P$. falciparum (grown to 5300 nodes with $\theta=(0.32,0.04,0.05)$, and subsampled according to $\rho)$. The green dashed line corresponds to proportional decay relative to the sampling fraction and is computed as follows. Denote the observed sampling fraction of the $P$. falciparum PIN data set with $\rho_{\mathcal{D}}=0.24$, and a mean simulated summary to an ensemble of simulated PINs at sampling fraction $\rho$ with $\overline{\mathcal{S}}_{\theta, \rho}$. Given another sampling fraction $\rho^{\prime}$, the slope of the green dashed line is $\left(\overline{\mathcal{S}}_{\theta, \rho_{\mathcal{D}}}-\frac{\rho^{\prime}}{\rho_{\mathcal{D}}} \overline{\mathcal{S}}_{\theta, \rho_{\mathcal{D}}}\right) /\left(\rho_{\mathcal{D}}-\rho^{\prime}\right)$. We considered $\rho^{\prime}=0.22$; the intercept is computed analogously. Apart from CC , all other summaries were distorted under decreasing $\rho$. While $\overline{\mathrm{ND}}$ decreased linear to the sampling fraction, PL decreased sublinearily and TRIA superlinearily. This indicates that at low sampling fraction, global aspects of PINs are much less distorted than motif counts in biological network data.

| Network Summaries | avg. acc. prob. | $\operatorname{GR}\left(\boldsymbol{\delta}_{\text {Div }}\right)$ | $\mathrm{GR}\left(\boldsymbol{\delta}_{\text {Att }}\right)$ | $\mathrm{GR}(\boldsymbol{\alpha})$ |
| :--- | :---: | :---: | :---: | :---: |
| ND | 0.15 | 1.00 | 1.01 | 1.00 |
| WR+DIA+CC+ | 0.02 | 1.00 | 1.00 | 1.00 |
| $\overline{\text { ND }+ \text { FRAG }}$ |  |  |  |  |
| WR+ND+ | 0.06 | 1.01 | 1.01 | 1.00 |
| CC+FRAG | 0.03 | 1.00 | 1.01 | 1.01 |
| ND+PL+ |  |  |  |  |
| DIA+CC+FRAG |  |  |  |  |

Table S.1: Convergence analysis of LFI on the H. pylori PIN data set. For each set of summaries, the average empirical acceptance probabilities and the GR-statistics are reported for each evolution parameter; see text in the Supporting Information. All runs converge well as indicated by the GR-statistic, and the empirical acceptance probabilities are between $[0.02,0.15]$ to facilitate cross-comparison.

## Convergence analysis of likelihood-free inference.

MCMC is guaranteed to converge in distribution to the posterior under some regularity conditions. The MCMC output may be directly used to determine the burn-in period, at which the Markov chain has not yet converged. After burn-in, MCMC continues and is taken to sample from the posterior. Such analysis is particularly important for likelihood-free inference within MCMC, because of (i) the nontrivial choice of $\varepsilon$ in combination with the set of summaries $\mathbb{S}$ and (ii) the nature of the ABC algorithm.

Suppose we start many Markov chains at overdispersed initial parameter values. A standard technique to detect that all chains have not yet converged is to compare the variance within one chain against the variance between all chains [41]: if this ratio exceeds 1.2 , then convergence is usually rejected. In this case we argue that the independent chains have not explored the state space sufficiently well, and must run longer.

Tables S. 1 and S. 2 document the empirical acceptance probabilities and the Gelman-Rubin test for all Markov chain runs presented in the main text. MCMC output is taken to represent posterior samples after 800 iterations if the Gelman-Rubin statistic does not exceed 1.2. This criterion ensures that posterior samples are taken after tempering the MCMC chains.

| Incompleteness | avg. acc. prob. | $\operatorname{GR}\left(\boldsymbol{\delta}_{\text {Div }}\right)$ | $\mathrm{GR}\left(\boldsymbol{\delta}_{\text {Att }}\right)$ | $\mathrm{GR}(\boldsymbol{\alpha})$ |
| :--- | :---: | :---: | :---: | :---: |
| $1271 / 5300$ | 0.04 | 1.00 | 1.00 | 1.00 |
| $900 / 5300$ | 0.03 | 1.00 | 1.03 | 1.02 |
| $600 / 5300$ | 0.02 | 1.08 | 1.03 | 1.06 |

Table S.2: Convergence analysis of LFI on the P. falciparum PIN data set. For PIN data sets of increasing incompleteness, we report the average empirical acceptance probabilities and the GR-statistic for the evolution parameters; see text in the Supporting Information. LFI with $\rho \geq 0.17$ (900/5300) converged very well as indicated by the GR-statistic. With increasing $\rho$, LFI converged less well.

## Preliminary Tests of LFI on simulated protein networks

Accuracy PIN test data sets were generated based on model (1) with the parameter $\theta=(0.4,0.3,0.3)$. To simulate incompleteness, we grew the PIN test data sets to 120 proteins and retained a subgraph of 100 proteins. For each test, 4 independent PIN test data sets were generated in this way. We believe that the performed tests were demanding because the protein test networks are (1) very small and (2) highly variable so that characteristics genuine to the evolution mechanism and yet common to all networks are highly diluted. Then we applied the LFI scheme detailed in Materials and Methods, with each Markov chain using a different PIN test data set. We tried a number of different sets of network summaries and studied the accuracy in re-estimating $\theta$. LFI based on $\mathrm{WR}+\mathrm{DIA}+\mathrm{CC}+\overline{\mathrm{ND}}+$ FRAG was particularly successful. For example, Figure S. 7 compares how accurately LFI based on WR + DIA $+\mathrm{CC}+\overline{\mathrm{ND}}+$ FRAG and LFI based on ND re-estimate $\theta$ in terms of 2D-histograms of the posterior of $\theta$.

Sensitivity For a number of different summaries, we also studied if LFI is sensitive to changes in the evolution model. Similar as above, we generated PIN test data sets according to DDa or PA with the parameters $\theta=$ $(0.4,0.3,0)$, and $\theta=(0.4,0.3,1)$ respectively. LFI based on $\mathrm{WR}+\mathrm{DIA}+\mathrm{CC}+\overline{\mathrm{ND}}+\mathrm{FRAG}$ may distinguish between DDa and PA. Figure S. 6 possibly illustrates best that the mixture parameter $\alpha$ is in favor of the correct model (orange), and that the marginal posterior of $\alpha$ separate clearly. Similar results were obtained for other tests based on 4 or more network summaries (data not shown).

In summary, our method correctly predicted the true model parameter with accuracy, converged quickly and clearly distinguished between different evolution mechanisms only when 4 or more different summaries with non-zero gradient and moderate variation were used. Notably, ND alone could not re-estimate the true model parameters confidently, while the set of summaries composed of WR + DIA $+\mathrm{CC}+\overline{\mathrm{ND}}+$ FRAG performed very well relative to other combinations. We could not reproduce similar results without either tempering or smoothing.


Figure S.6: Violin plot of the marginal posterior of $\alpha$ for LFI based on $\mathrm{WR}+\mathrm{DIA}+\mathrm{CC}+\overline{\mathrm{ND}}+$ FRAG on PIN test data sets. (left) The PIN test data sets were generated by $\theta=(0.4,0.3,0)$. (right) The PIN test data sets were generated by $\theta=(0.4,0.3,1)$. In both cases, the re-estimated marginal posterior of $\alpha$ is in favor of the correct model (orange lines), and the bulk of the marginal posteriors are clearly separated.


Figure S.7: LFI to re-estimate the parameter from 4 randomly created test PINs generated with $\theta=(0.4,0.3,0.3)$ as described in the Supplementary Information. (a,c,e) 2D-histograms of the posterior parameters with LFI based on $\mathrm{WR}+\mathrm{DIA}+\mathrm{CC}+\overline{\mathrm{ND}}+\mathrm{FRAG}$. The posteriors are centered around the true parameter (red). (b,d,f) For comparison, 2D-histograms of the posterior parameters based on LFI with ND. Here, the posteriors appear diffuse and are not centered around the true parameter.

| Network Summaries | avg. acc. prob. | $\operatorname{GR}\left(\boldsymbol{\delta}_{\text {Div }}\right)$ | $\operatorname{GR}\left(\boldsymbol{\delta}_{\text {Att }}\right)$ | $\operatorname{GR}(\boldsymbol{\alpha})$ |
| :--- | :---: | :---: | :---: | :---: |
| ND | 0.30 | 1.02 | 1.01 | 1.01 |
| WR+DIA + CC + | 0.31 | 1.03 | 1.04 | 1.10 |
| $\overline{\text { ND }}+$ FRAG |  |  |  |  |

Table S.3: Convergence analysis for LFI to fit the PIN test data sets, accompanying Figure S.7. For each set of summaries, the average empirical acceptance probabilities and the GR-statistics are reported for each evolution parameter; see text in the Supporting Information. All runs converge as indicated by GR, and the empirical acceptance probabilities are similar, allowing a cross-comparison.

