communications physics

https://doi.org/10.1038/s42005-022-00846-x OPEN

Author Correction: Hypergraph reconstruction from network data

Jean-Gabriel Young, Giovanni Petri & Tiago P. Peixoto

Correction to: *Communications Physics* https://doi.org/10.1038/s42005-021-00637-w, published online 15 June 2021.

It has recently come to the authors' attention that the marginal distribution introduced in Eq. (12) contains a mistake and is inconsistent with the derivation given for the empirical prior in Eqs. (10)–(11).

The authors have corrected the mistake and verified that the properties of Eq. (12) derived in the original article still hold. They have also re-ran every calculation using the updated equation.

The empirical findings are qualitatively unchanged; the authors have found highly similar compression ratios when comparing the clique descriptions to the optimal descriptions obtained by minimising the description length (DL). The exact values of the minimum description length (MDL) are different since the modification to Eq. (12) changes the absolute value of $P(H)$, and since the Monte Carlo Markov chain algorithm is stochastic and therefore yields slightly different results from runs to runs. Below is a detailed list of changes brought about by modifying Eq. (12) and running new simulations.

Changes to the main text

In the subsection "Hypergraph prior", in the 7th paragraph right before Eq. (10): "calculated" has been replaced by "approximated."

The original Eq. (10) read

$$\langle E(\boldsymbol{\nu})\rangle = \binom{N}{2}\left[1-\prod_{k=2}^{L}\left(e^{-\nu_k}\right)^{\binom{N}{k-1}}\right],$$

and has now been replaced by

$$\langle E(\boldsymbol{\nu})\rangle = \sum_k \nu_k \binom{N}{k}.$$

The justification for Eq. (10): "first computing the reciprocal of the probability that two nodes are not connected by any hyperedge in the hypergraph and then multiplying the result by the total number of node pairs" in the original text has been replaced by: "by assuming that hyperedges do not overlap on average."

The original Eq. (11) read

$$\mu = (L-1)\log\left(\frac{1}{1-E/\binom{N}{2}}\right),$$

and has now been replaced by:

$$\mu = E/(L-1).$$

Eq. (12) was originally

$$P(H) = \prod_{k=2}^{L}\frac{E_k!}{Z_k\binom{N}{k}^{E_k}\mu}\left[\frac{N-k+1}{k}+\frac{1}{\mu}\right]^{-(E_k+1)},$$

and has now been replaced by:

$$P(H)=\prod_{k=2}^{L}\frac{E_k!}{Z_k\mu\binom{N}{k}^{E_k}}\left[\frac{1}{\mu}+1\right]^{-(E_k+1)}.$$

The text immediately following Eq. (12) which read "We note that μ diverges as the density $E/\binom{N}{2}$ of G approaches one, correctly reflecting the fact that even an infinitely dense hypergraph could have generated the data. This divergence is a sign that our empirical prior is not well-defined in the extremely dense limit. But as we have discussed in the introduction, the empirical networks we typically encounter are sparse by construction—we need not worry about this limit in practice.", has been removed since the approximated equation for μ does not diverge. Instead, Eq. (12) is now followed by "which is the equation we will use henceforth, with $\mu = E/(L-1)$."

The scaling equation for $\log P(H|G)$ appearing in "Results and discussion", under the heading "Properties of the posterior distribution," was

$$\log P(H|G)\sim-\alpha(1+\beta)\binom{N}{k}\log\left(\frac{\frac{N+k-1}{k}+\frac{1}{\mu}}{\alpha}\right),$$

and now reads:

$$\log P(H)\sim-\alpha(1+\beta)\binom{N}{k}\log\left[\frac{1+\frac{1}{\mu}}{\alpha}\right].$$

The conditioning of this equation on G is discussed in the text following the equation, which originally stated:

"This equation tells us that the log-posterior $\log P(H|G)$ decreases with growing β, because the argument of the logarithm is at least one. Furthermore we have $P(G|H) = 1$ by construction, ... "

It now reads: "This equation tells us that the log-posterior $\log P(H|G)$ decreases with growing β, because the argument of the logarithm is greater or equal to one. Furthermore, the likelihood equals one by construction...."

The ratio of posterior distributions under changes to a minimal hypergraph originally appearing in the third paragraph of the Results subsection titled "Properties of the posterior distribution" was written as

$$\frac{P(H'_m|G)}{P(H_m|G)}=\frac{E_k+1}{\binom{N}{k}\left(\frac{N+k-1}{k}+\frac{1}{\mu}\right)},$$

and has now been corrected to:

$$\frac{P(H'_m|G)}{P(H_m|G)}=\frac{Z_k}{Z'_k}\frac{E_k+1}{\binom{N}{k}}\left[\frac{1}{\mu}+1\right]^{-1}.$$

The paragraph following this equation read "This ratio is smaller than one: the minimal property of H_m implies that $E_k<\binom{N}{2}$, and the term in the parenthesis is greater than one because $N\gg k$. As a result, adding a spurious hyperedge to a minimal hypergraph decreases the posterior probability. As a corollary of the two above observations, we conclude that the minimal hypergraphs are high-quality local maxima of $P(H|G)$. We cannot simply pick one of these optima as our reconstruction, however, because there may exist multiple ones of comparable quality. Further, non-optimal hypergraphs may account for a significant fraction of the posterior probability in principle. Instead, we handle these possibly conflicting descriptions by combining them."

It now reads "where Z'_k is the quantity in Eq. (7) for the modified minimal hypergraph, and Z_k is the same quantity for the minimal hypergraph. One can show that this ratio is always smaller than one and that, as a result, adding a spurious hyperedge to a minimal hypergraph decreases the posterior probability. The proof is straightforward and relies on the observation that for a minimal hypergraph, we have $E_k\le\binom{N}{2}$, $Z_k = 1$, and $Z'_k = 1$ or $Z'_k = 2$. The result follows by direct computation when $E_k<\binom{N}{2}$ and uses the fact that that $Z'_k = 2$ when $E_k=\binom{N}{2}$ (because adding a single hyperedge to a completely connected minimal hypergraph means one has to double-up one hyperedge)."

In 'Results and discussion", under the heading "Detailed case study of higher-order interactions in an empirical network," the number of nodes and edges in the Football dataset have been added.

The text has changed from "The nodes of this network represent teams playing..." to "The 115 nodes of this network represent teams playing..." and "The relationships between teams are viewed through the lens of pairwise interactions ..." now reads "The relationships between teams are viewed through the lens of 613 pairwise relationships…"

Under the subheading "Best model fit," four numerical changes have been made: The hyperedges of H^* involving more than two nodes were 86, and now are 30. The description length was 4123.3 bits and now is 2405.8 bits, which represents a 43.3% saving (instead of the previous 33.6%) over the description length of the maximal clique hypergraph, which now is of 4246.5 bits (instead of 6208.5 bits).

On the same page, under the "Probabilistic descriptions" subheading before Eq. (17), the uncertain range was $[\alpha/2, 1-\alpha/2]$ and is now described as $[\alpha, 1-\alpha]$.

This change appears in two locations: before Eq. (17), and in the caption of Fig. 6.

The following sentence read "With a threshold of $\alpha = 0.05$, we find six uncertain triangles (hyperedges on three nodes) and five uncertain edges in the Football data", and now is instead "With a threshold of $\alpha = 0.05$, we find 16 uncertain triangles (hyperedges on three nodes), 70 uncertain edges, and 9 additional uncertain interactions of higher orders."

After Eq. (17), the threshold for uncertainty is now updated to $S^* \approx 0.286$ (instead of $S^* = 0.169$).

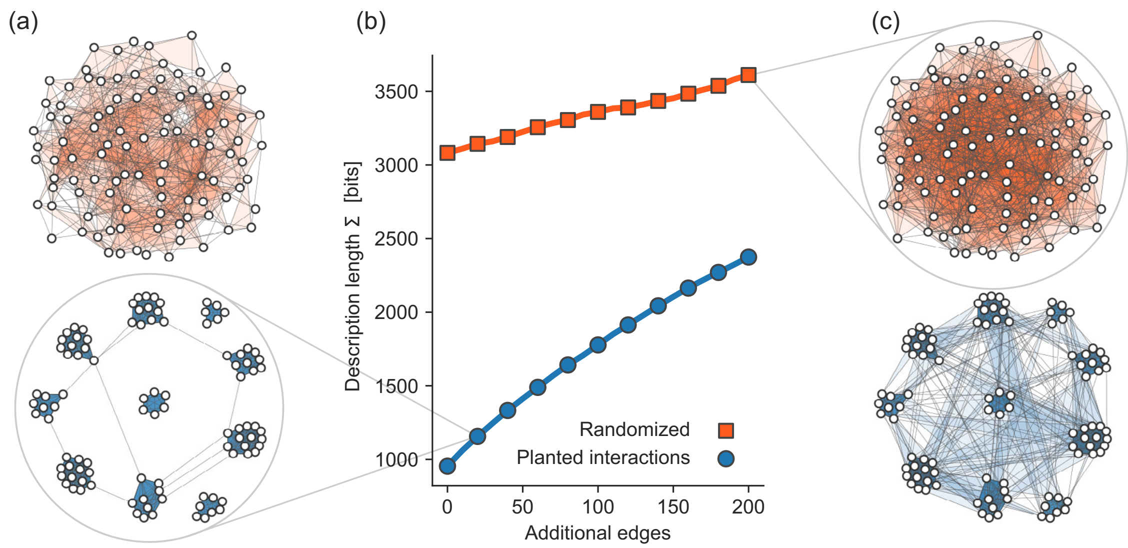
On page 9, first paragraph, the correlation coefficients reported under the heading "Systematic analysis of higher-order interactions in empirical networks" have been updated to reflect the results obtained with the new posterior distribution. The average degree of the nodes still correlates with compression ($\tau = 0.52$, previously 0.53) and, as before, the average local clustering does not correlate with compression ($\tau = 0.03$, previously $\tau = -0.07$). The average interaction size is no longer significantly correlated with the tested properties. As a consequence, in the second paragraph, the sentence "The correlation between local properties and interaction size is not as strong as with compression, but there are some dependencies ($\tau = 0.40$ and $\tau = 0.27$ for the degree and local clustering, respectively). These might be partly explained by constraints on the possible values that the average interaction size $\langle s \rangle$ can adopt." is now rewritten as: "The correlation between local properties and interaction size is weak ($\tau = 0.09$ and $\tau = 0.12$ for the degree and local clustering, respectively). Nonetheless, we expect some weak dependencies as these network properties put constraints on the possible values that the average interaction size $\langle s \rangle$ can adopt."

In the second paragraph of page 9, the average interaction size has been updated. The sentence "Other datasets yield hypergraphs with large interactions on average, involving as many as five nodes in the airport network." is now changed to "Other datasets yield hypergraphs with large interactions on average, involving as many as 4.4 nodes in the airport network."

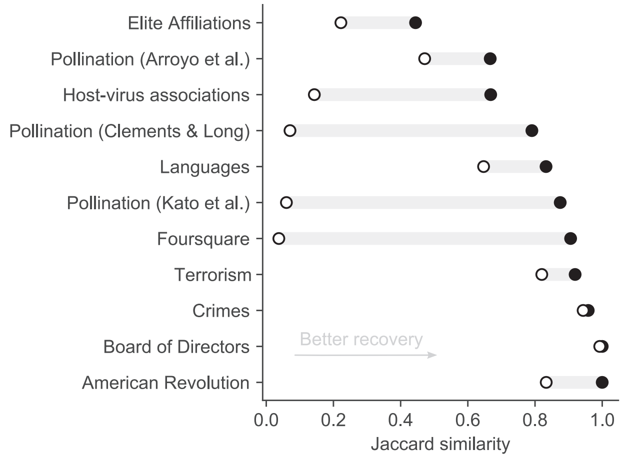
In the third paragraph of the conclusion, at page 10, the word "undoubtedly" was removed from the sentence "The method we have proposed here is undoubtedly one of the simplest instantiations…", and now reads "The method we have proposed here is one of the simplest instantiations…"

Changes to the figures

Figure 3 has been replaced to reflect the new numerical value of $P(H|G)$. In panel (b), the qualitative behavior of the minimum description length (MDL) as a function of the density remains identical although the Y axis values have changed from the range [≈ 1000–3500] to the range [500–3000]. In panel (a) and (c), fewer triangles are found by the method in the randomized case, shown in orange. See Correction Fig. 1 for the original version of Fig. 3.



Correction Fig. 1 Original version of Figure 3.



Correction Fig. 2 Original version of Figure 4.

Figure 4 has been replaced to reflect the results obtained with the new posterior distribution. The results are identical, except for the Jaccard coefficient of the MDL reconstruction (filled symbol) for the following datasets: "Elite affiliation" (previous value 0.44 ⟶ updated value 0.25) "Pollination (Arroyo et al.)" (previous value 0.67 ⟶ updated value 0.66), "Pollination (Clements & Long)" (previous value 0.79 ⟶ updated value 0.81), and "Foursquare" (previous value 0.91 ⟶ updated value 0.90). See Correction Fig. 2 for the original version of Fig. 4.

Figure 5 has been replaced to reflect the results obtained with the new posterior distribution. Panel (a) is now reported on a log scale. As a result of updating the distribution, there no longer are any triangles in the best fit and hence the bar for hyperedges of size 3 is now absent, while it originally showed roughly 40 hyperedges. Accordingly, there are no longer any triangles highlighted in the new panel (b). See Correction Fig. 3 for the original version of Fig. 5.

Figure 6 has been replaced to reflect the results obtained with the new posterior distribution. Panel (a) shows more uncertain edges, and the natural gap in uncertainty now occurs at $S^* \approx 0.286$ (instead of the value $S^* = 0.169$ reported in the original paper). Panel (b) shows the location of the new uncertain edges and triangles, which number increases with respect to the original version and as reported in the updated version of the main text.

The following sentence has been added to the caption of Fig. 6: "Remaining uncertain interactions are shown in gray." The uncertain range was $[\alpha/2, 1-\alpha/2]$ and is now described as $[\alpha, 1-\alpha]$.

The new simulations were performed with 4000 samples instead of 2000, and are separated by 1000 sweeps instead of 2000. Accordingly, the last sentence of the caption of Fig. 6 was "All results are computed with 2,000 Monte Carlo samples from the posterior distribution each separated by 2,000 complete sweep of the factor graph.". Now it reads: "All results are computed with 4,000 Monte Carlo samples from the posterior distribution each separated by 1,000 complete sweep of the factor graph." See Correction Fig. 4 for the original version of Fig. 6.

Figure 7 has been replaced to reflect the results obtained with the new posterior distribution. The results and data trends are similar to the ones pictured in the original version of Fig. 7, though the exact numerical values have changed (see the correction to Supplementary Table 2 below for detailed numerical values). See Correction Fig. 5 for the original version of Fig. 7.

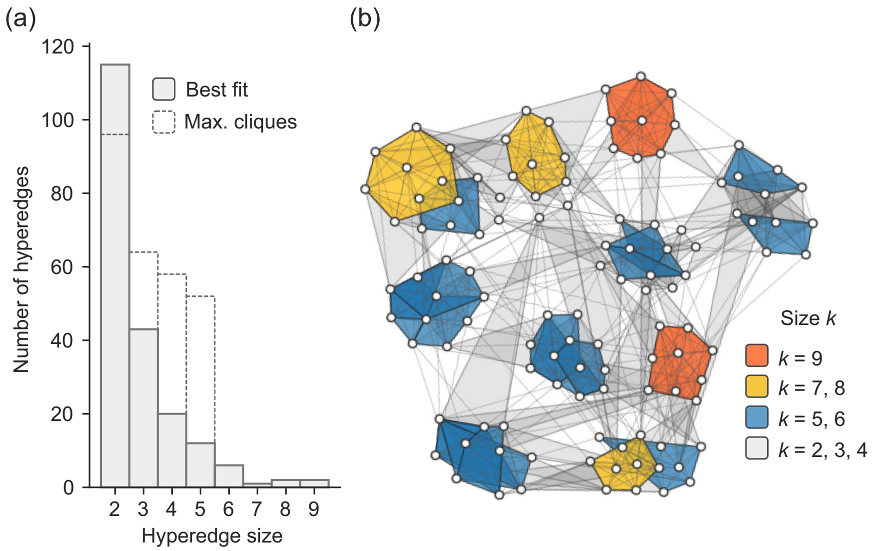
Changes to the supplementary information

Supplementary note 1. Equations (2) and (4) have been updated to reflect the matching changes made to equations (12) and (11) of the main text, respectively. In particular, Supplementary Equation (2) was

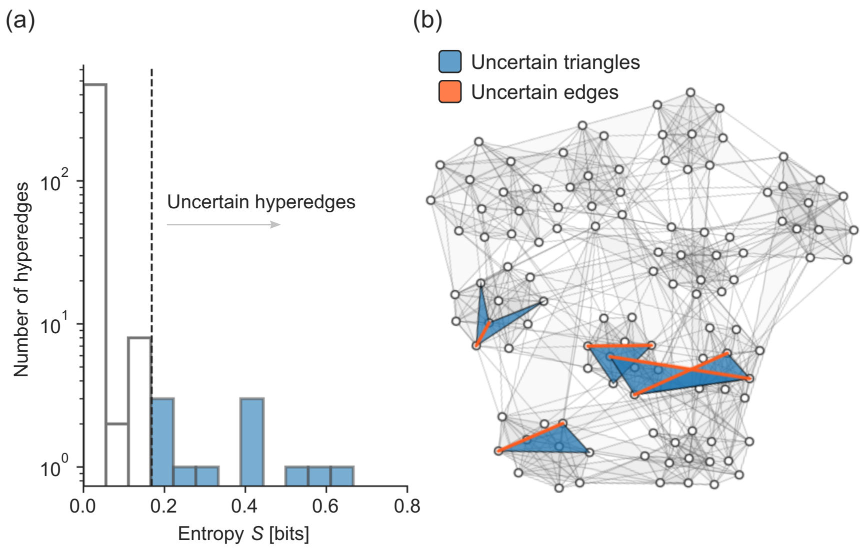
$$P(H) = \prod_{k=2}^{L} \frac{E_k!}{Z_k \binom{N}{k}^{E_k} \mu} \left[\frac{N-k+1}{k} + \frac{1}{\mu}\right]^{-(E_k+1)}.$$

And has now been replaced by:

$$P(H) = \prod_{k=2}^{L} \frac{E_k!}{Z_k \mu \binom{N}{k}^{E_k}} \left[\frac{1}{\mu} + 1\right]^{-(E_k+1)}.$$



Correction Fig. 3 Original version of Figure 5.



Correction Fig. 4 Original version of Figure 6.

Supplementary Equation (4) was

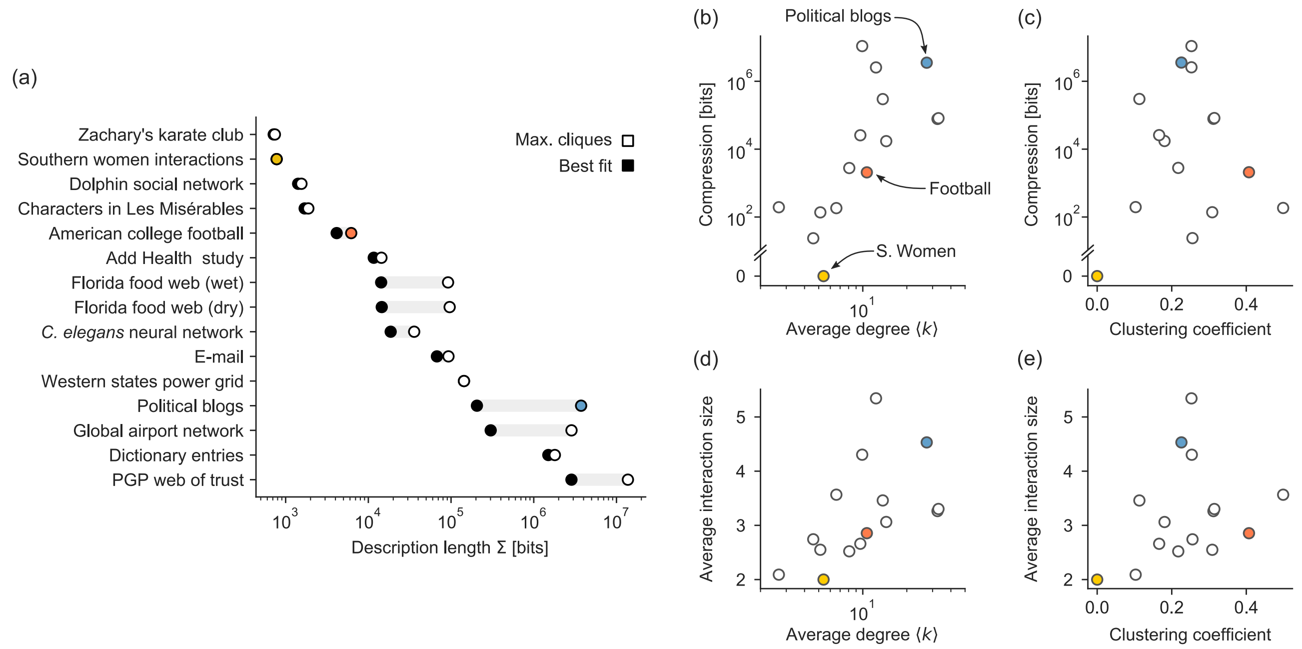
$$\mu = (L-1)\log\left(\frac{1}{1-E/\binom{N}{2}}\right),$$

And has now been replaced by:

$$\mu = E/(L-1)$$

Supplementary Table 1: The column "J_{MDL}" of the Supplementary Table 1 has been updated with new Jaccard coefficients. It changed for the four datasets, as follows: "Elite affiliation" (previous value 0.44 → updated value 0.25) "Pollination (Arroyo et al.)" (previous value 0.67 → updated value 0.66), "Pollination (Clements & Long)" (previous value 0.79 → updated value 0.81), and "Foursquare" (previous value 0.91 → updated value 0.90).

The caption title of Supplementary Table 1 was "Properties of the empirical bipartite networks analyzed in Section II.E of the main text." which erroneously referred to an invalid reference (Section II.E) and has now been changed to "Properties of the empirical bipartite networks analyzed in Fig. 4 of the main text."



Correction Fig. 5 Original version of Figure 7.

Dataset	E	N	$\langle k \rangle$	C	DL [bits]	Clique DL [bits]	$\langle s \rangle$
Zachary's karate club [14]	78	34	4.6	0.26	717	740	2.74
Southern women interactions [15]	89	32	5.5	0.00	778	778	2.00
Dolphin social network [16]	159	62	5.1	0.31	1,416	1,553	2.55
Characters in Les Misérables [17]	254	77	6.6	0.50	1,698	1,881	3.57
American college football [18]	613	115	10.7	0.41	4,139	6,208	2.86
Add Health study [19]	1,136	280	8.1	0.22	11,618	14,396	2.52
Florida food web (wet) [20]	2,075	128	32.4	0.31	14,281	91,996	3.26
Florida food web (dry) [20]	2,106	128	32.9	0.31	14,521	96,200	3.30
C. elegans neural network [21]	2,148	297	14.5	0.18	18,619	35,829	3.06
E-mail [22]	5,451	1,136	9.6	0.17	67,433	93,088	2.66
Western states power grid [23]	6,594	4,941	2.7	0.10	143,233	143,427	2.09
Political blogs [24]	16,714	1,220	27.4	0.23	205,489	3,737,574	4.53
Global airport network [25]	20,232	3,290	12.3	0.25	301,459	2,860,031	5.34
Dictionary entries [26]	91,471	13,353	13.7	0.11	1,499,765	1,798,071	3.46
PGP web of trust [27]	197,150	39,828	9.9	0.25	2,855,554	13,753,620	4.30

Correction Fig. 6 Original version of Supplementary Table 2.

Supplementary Table 2: Al the values found in the columns "DL [bits]," "Clique DL [bits]," and "⟨s⟩" have been updated according to the new numerical simulations. The DL values are all lower, generally with a steeper decrease in DL than Clique DL. The average interaction size ⟨s⟩ is also smaller on average in the updated table.

The caption title of Supplementary Table 2 was "Properties of the empirical bipartite networks analyzed in Section II.G of the main text." which erroneously referred to an invalid reference (Section II.G) and has now been changed to "Properties of the empirical bipartite networks analyzed in Fig. 7 of the main text." See Correction Fig. 6 for the original version of Supplementary Table 2.

All corrections described above have now been implemented in both the HTML and PDF version of the article. The Supplementary Information file has also been updated with the corrected version.

Published online: 29 March 2022

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42005-022-00846-x.

Acknowledgements

The authors thank Huan Wang, Yanbang Wang, and Kyuhan Lee for bringing the mistake in Eq. (12) to their attention.