Network reconstruction from infection cascades
Supplementary Information

Alfredo Braunstein*
DISAT, Politecnico di Torino, Corso Duca Degli Abruzzi 24, 10129 Torino, Italian Institute for Genomic Medicine, Via Nizza 52, 10124 Torino, Collegio Carlo Alberto, Piazza Arbarello 8, 10122 Torino & INFN Sezione di Torino, Via P. Giuria 1, I-10125 Torino, Italy

Alessandro Ingrosso†
Center for Theoretical Neuroscience, Columbia University, New York, USA

Anna Paola Muntoni‡
Laboratoire de physique théorique, Département de physique de l'ENS, École normale supérieure, PSL University, Sorbonne Université, CNRS, 75005 Paris, France and DISAT, Politecnico di Torino, Corso Duca Degli Abruzzi 24, 10129 Torino

I. BP EQUATION: EFFICIENT DISENTANGLED IMPLEMENTATION

We would like to use a factor graph representation that maintains the same topological properties of the original graph of contacts, in order to guarantee that BP is exact when the original contact graph is a tree. Following an approach developed in previous works [1–3], we proceed to disentangle the factor graph by grouping pairs of infection times (t_i, t_j) in the same variable node. For convenience, we will keep all variable nodes $\{t_i\}$ but we will also introduce for each edge (i, j) emerging from a node i a set of copies $t_i^{(j)}$ of the infection time t_i, that will be forced to take the common value t_i by including the constraint $\prod_{k\in\partial i}\delta(t_i^{(k)}, t_i)$ in an additional factor ϕ_i.

The factors ϕ_i depend on infection times and transmission delays just through the sums $t_i^{(j)} + s_{ij}$, so that it is more convenient to introduce the variables $t_{ij} = t_i^{(j)} + s_{ij}$ and express the dependencies through the pairs $(t_i^{(j)}, t_{ij})$.

Finally it is convenient to group the variable g_i with the corresponding infection times t_i in the same variable node, replace g_i and g_j by their copies $g_i^{(j)}$ and $g_j^{(i)}$ in the edge constraints $\omega_{ij}(t_{ij} - t_i^{(j)}|g_i^{(i)})$ and $\omega_{ji}(t_{ji} - t_j|g_j^{(i)})$ and impose the identity $\prod_{k\in\partial i}\delta(g_i^{(k)}, g_i)$ for each node i. The resulting disentangled factor graph appears in Fig. 1.

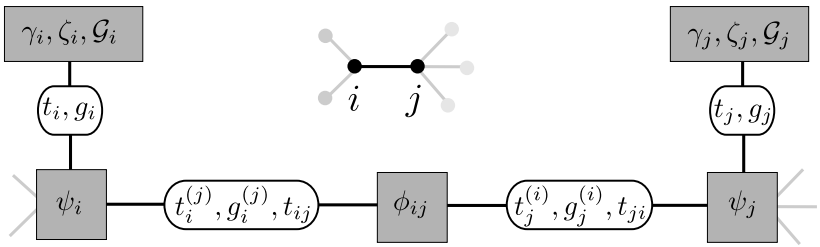

Figure 1. Disentangled Factor graph representation of the graphical model. White round nodes correspond to variables, gray rectangle nodes correspond to factors (or constraints). The topology of the disentangled factor graph follows the one of the original contact network.

* alfredo.braunstein@polito.it
† ai2367@columbia.edu
‡ anna.muntoni@polito.it

An efficient form for the update equations of the ψ_i factor nodes is the following:

$$p_{\psi_i\to j}\left(t_i^{(j)},t_{ji},g_i^{(j)}\right)\propto\sum_{g_i,t_i}\sum_{\left\{t_i^{(k)},t_{ki},g_i^{(k)}\right\}} m_{i\to\psi_i}\left(t_i,g_i\right)\times \tag{1}$$

$$\times\prod_{k\in\partial i\setminus j} m_{k\to\psi_i}\left(t_i^{(k)},t_{ki},g_i^{(k)}\right)\psi_i\left(t_i,g_i,\left\{\left(t_i^{(k)},t_{ki},g_i^{(k)}\right)\right\}_{k\in\partial i}\right)$$

$$\propto m_{i\to\psi_i}\left(t_i^{(j)},g_i^{(j)}\right)\sum_{t_{ki}}\prod_{k\in\partial i\setminus j} m_{k\to\psi_i}\left(t_i^{(j)},t_{ki},g_i^{(j)}\right)\times \tag{2}$$

$$\times\left[\delta\left(t_i^{(j)},0\right)+\delta\left(t_i^{(j)},\left(1+\min_{k\in\partial i}\left\{t_{ki}\right\}\right)\right)\right]$$

$$\propto\delta\left(t_i^{(j)},0\right)m_{i\to\psi_i}\left(0,g_i^{(j)}\right)\prod_{k\in\partial i\setminus j}\sum_{t_{ki}} m_{k\to\psi_i}\left(0,t_{ki},g_i^{(j)}\right)+ \tag{3}$$

$$+ m_{i\to\psi_i}\left(t_i^{(j)},g_i^{(j)}\right)\mathbb{I}\left(t_i^{(j)}\le t_{ji}+1\right)\prod_{k\in\partial i\setminus j}\sum_{t_{ki}\ge t_i^{(j)}-1} m_{k\to\psi_i}\left(t_i^{(j)},t_{ki},g_i^{(j)}\right)$$

$$- m_{i\to\psi_i}\left(t_i^{(j)},g_i^{(j)}\right)\mathbb{I}\left(t_i^{(j)}< t_{ji}+1\right)\prod_{k\in\partial i\setminus j}\sum_{t_{ki}> t_i^{(j)}-1} m_{k\to\psi_i}\left(t_i^{(j)},t_{ki},g_i^{(j)}\right)$$

where in (3) we use the fact that

$$\delta\left(t_i,\left(1+\min_{j\in\partial i}\left\{t_{ji}\right\}\right)\right)=\prod_{j\in\partial i}\mathbb{I}\left(t_i\le t_{ji}+1\right)-\prod_{j\in\partial i}\mathbb{I}\left(t_i< t_{ji}+1\right).$$

Up to now, messages depend on the T^2G values $\left(t_i^{(k)},t_{ki},g_i^{(k)}\right)$. It is however possible to use more concise representation, retaining just information on the relative timing between infection time $t_i^{(j)}$ for a node i and the infection propagation time t_{ji} on its link with node j, introducing the variables

$$\sigma_{ji}=1+\text{sign}\left(t_{ji}-\left(t_i^{(j)}-1\right)\right), \tag{4}$$

In order to switch to the simplified representation with $(\sigma_{ji},\sigma_{ij})$ variables defined in (4) instead of t_{ji},t_{ij} ones, we will proceed as follows. In equation (3) we can easily group the sums over different configurations of $\left(t_{ki},t_i^{(j)}\right)$ and write:

$$p_{\psi_i\to j}\left(t_i^{(j)},\sigma_{ji},g_i^{(j)}\right)\propto\delta\left(t_i^{(j)},0\right)m_{i\to\psi_i}\left(0,g_i^{(j)}\right)\prod_{k\in\partial i\setminus j}\sum_{\sigma_{ki}} m_{k\to\psi_i}\left(0,\sigma_{ki},g_i^{(j)}\right)+ \tag{5}$$

$$+ m_{i\to\psi_i}\left(t_i^{(j)},g_i^{(j)}\right)\mathbb{I}\left(\sigma_{ji}=1,2\right)\prod_{k\in\partial i\setminus j}\sum_{\sigma_{ki}=1,2} m_{k\to\psi_i}\left(t_i^{(j)},\sigma_{ki},g_i^{(j)}\right)$$

$$- m_{i\to\psi_i}\left(t_i^{(j)},g_i^{(j)}\right)\mathbb{I}\left(\sigma_{ji}=2\right)\prod_{k\in\partial i\setminus j} m_{k\to\psi_i}\left(t_i^{(j)},2,g_i^{(j)}\right)$$

Similarly, the outgoing message to the (t_i,g_i) variable node is:

$$p_{\psi_i\to i}\left(t_i,g_i\right)\propto\delta\left(t_i,0\right)\prod_{k\in\partial i}\sum_{\sigma_{ki}} m_{k\to\psi_i}\left(0,\sigma_{ki},g_i\right)+ \tag{6}$$

$$+\prod_{k\in\partial i}\sum_{\sigma_{ki}=1,2} m_{k\to\psi_i}\left(t_i,\sigma_{ki},g_i\right)$$

$$-\prod_{k\in\partial i} m_{k\to\psi_i}\left(t_i,2,g_i\right)$$

In the simplified (t,σ,g) representation for the messages, the update equation for the ϕ_{ij} nodes reads:

$$p_{\phi_{ij}\to j}\left(t_j,\sigma_{ij},g_j\right)\propto\sum_{t_i,\sigma_{ji},g_i}\Omega\left(t_i,t_j,\sigma_{ij},\sigma_{ji},g_i,g_j\right)m_{i\to\phi_{ij}}\left(t_i,\sigma_{ji},g_i\right) \tag{7}$$

where:

$$\Omega\left(t_i,t_j,\sigma_{ij},\sigma_{ji},g_i,g_j\right)=\begin{cases}\chi\left(t_i,t_j,\sigma_{ij},g_i\right) & :t_i<t_j,\ \sigma_{ji}=2,\ \sigma_{ij}\neq 2\\ \chi\left(t_i,t_j,\sigma_{ij},g_i\right)+(1-\lambda)^{g_i+1} & :t_i<t_j,\ \sigma_{ji}=2,\ \sigma_{ji}=2\\ \chi\left(t_j,t_i,\sigma_{ji},g_j\right) & :t_j<t_i,\ \sigma_{ji}=2,\ \sigma_{ji}\neq 2\\ \chi\left(t_j,t_i,\sigma_{ji},g_j\right)+(1-\lambda)^{g_j+1} & :t_j<t_i,\ \sigma_{ij}=2,\ \sigma_{ji}=2\\ 1 & :t_i=t_j,\ \sigma_{ji}=\sigma_{ij}=2\\ 0 & :otherwise\end{cases} \tag{8}$$

and

$$\chi\left(t_1,t_2,\sigma,g\right)=\sum_{t=t_1}^{t_1+g}\delta\left(\sigma\left(t_2,t\right),\sigma\right)\lambda\left(1-\lambda\right)^{t-t_1} \tag{9}$$

Simple algebra and precalculation of terms in (7)-(9) brings a significant optimization for updates involving the factor node ϕ_{ij} down to $O(T\,|E|)$ operations per update.

II. GRADIENT DESCENT UPDATES

The log-likelihood of the epidemic parameters is nothing but the (negated) free energy of the model. In the Bethe approximation, it can be expressed as a sum of local terms which only depends on the BP messages:

$$-f=\sum_a f_a+\sum_i f_i-\sum_{(ia)} f_{(ia)} \tag{10}$$

where

$$f_a=\log\left(\sum_{\{z_i:i\in\partial a\}}F_a\left(\{z_i\}_{i\in\partial a}\right)\prod_{i\in\partial a}m_{i\to a}(z_i)\right) \tag{11}$$

$$f_{(ia)}=\log\left(\sum_{z_i}m_{i\to a}(z_i)p_{F_a\to i}(z_i)\right) \tag{12}$$

$$f_i=\log\left(\sum_{z_i}\prod_{b\in\partial i}p_{F_b\to i}(z_i)\right) \tag{13}$$

Since f is a function of all the BP messages, one would argue that this messages depend on the model parameters too, at every step in the BP algorithm. Actually, there is no need to consider this implicit $\{\lambda_{ij},\mu_i\}$ dependence if BP has reached its fixed point, that is when BP equations are satisfied and the messages are nothing else but Lagrange multipliers with respect to the constraint minimization of the Bethe free energy functional [4]. In the present parametrization, the only explicit dependence of free energy on epidemic parameters is in the factor node terms f_a's involving the compatibility functions $\phi_{ij}=\omega_{ij}\left(t_{ij}-t_i|g_i\right)\omega_{ji}\left(t_{ji}-t_j|g_j\right)$ and $\mathcal{G}_i\left(g_i\right)=\mu_i\left(1-\mu_i\right)^{g_i}$, and the gradient can be computed very easily. Please note that formulas below show the derivative of the free energy $f=-\mathcal{L}$: the GA updates of the log-likelihood only differ up to a minus sign. For the ϕ_{ij} nodes we have:

$$\frac{\partial f_{\phi_{ij}}}{\partial\lambda_{ij}}=\frac{\sum_{t_i,t_{ji},g_i,t_j,t_{ij},g_j}\frac{\partial\phi_{ij}}{\partial\lambda_{ij}}\left(t_i,t_{ji},g_i,t_j,t_{ij},g_j\right)m_{i\to\phi_{ij}}\left(t_i,t_{ji},g_i\right)m_{j\to\phi_{ij}}\left(t_j,t_{ij},g_j\right)}{\sum_{t_i,t_{ji},g_i,t_j,t_{ij},g_j}\phi_{ij}\left(t_i,t_{ji},g_i,t_j,t_{ij},g_j\right)m_{i\to\phi_{ij}}\left(t_i,t_{ji},g_i\right)m_{j\to\phi_{ij}}\left(t_j,t_{ij},g_j\right)} \tag{14}$$

where

$$\frac{\partial\phi_{ij}}{\partial\lambda_{ij}}=\begin{cases}1 & t_i<t_j\text{ and }t_i=t_{ij}<t_i+g_i\\ -\left(g_i-t_i\right)\lambda_{ij}\left(1-\lambda_{ij}\right)^{g_i-t_i-1} & t_i<t_j\text{ and }t_i<t_{ij}=t_i+g_i\\ \left(1-\lambda_{ij}\right)^{t_{ij}-t_i}-\left(t_{ij}-t_i\right)\lambda_{ij}\left(1-\lambda_{ij}\right)^{t_{ij}-t_i-1} & t_i<t_j\text{ and }t_i<t_{ij}<t_i+g_i\\ 1 & t_j<t_i\text{ and }t_j=t_j<t_j+g_j\\ -\left(g_j-t_j\right)\lambda_{ij}\left(1-\lambda_{ij}\right)^{g_j-t_j-1} & t_j<t_i\text{ and }t_j<t_{ji}=t_j+g_j\\ \left(1-\lambda_{ij}\right)^{t_{ji}-t_j}-\left(t_{ji}-t_j\right)\lambda_{ij}\left(1-\lambda_{ij}\right)^{t_{ji}-t_j-1} & t_j<t_i\text{ and }t_j<t_{ji}<t_j+g_j\\ 0 & \text{else}\end{cases} \tag{15}$$

In the simplified (t, σ, g) representation for the messages, equation (15) takes the form:

$$\frac{\partial \phi_{ij}}{\partial \lambda_{ij}} = \begin{cases} \chi\left(t_i, t_j, \sigma_{ij}, g_i\right) & t_i < t_j, \sigma_{ji} = 2, \sigma_{ij} \neq 2 \\ \chi\left(t_i, t_j, \sigma_{ij}, g_i\right) - \left(g_i + 1\right)\left(1 - \lambda\right)^{g_i} & t_i < t_j, \sigma_{ji} = 2, \sigma_{ij} = 2 \\ \chi\left(t_j, t_i, \sigma_{ji}, g_j\right) & t_j < t_i, \sigma_{ji} = 2, \sigma_{ij} \neq 2 \\ \chi\left(t_j, t_i, \sigma_{ji}, g_j\right) - \left(g_j + 1\right)\left(1 - \lambda\right)^{g_j} & t_j < t_i, \sigma_{ji} = 2, \sigma_{ij} = 2 \\ 0 & otherwise \end{cases} \tag{16}$$

where:

$$\chi\left(t_1, t_2, \sigma, g\right) = \sum_{t=t_1}^{t_1+g} \delta\left(\sigma\left(t_2, t\right), \sigma\right)\left(1 - \lambda_{ij}\right)^{t-t_1} - \left(t - t_1\right)\lambda_{ij}\left(1 - \lambda_{ij}\right)^{t-t_1-1} \tag{17}$$

For the $\mathcal{G}_i$ nodes we have:

$$\frac{\partial f_{\mathcal{G}_i}}{\partial \mu_i} = \frac{\sum_{g_i} \tilde{\mathcal{G}}_i(g_i) m_{i \to \mathcal{G}_i}(g_i)}{\sum_{g_i} \mathcal{G}_i(g_i) m_{i \to \mathcal{G}_i}(g_i)} \tag{18}$$

where

$$\tilde{\mathcal{G}}_i(g_i) = \begin{cases} \left(1 - \mu_i\right)^{g_i} - g_i \mu_i \left(1 - \mu_i\right)^{g_i - 1} & : g_i < G \\ G - G\left(1 - \mu_i\right)^{G-1} & : g_i = G. \end{cases} \tag{19}$$

III. ADDITIONAL RECONSTRUCTED INFLUENCE NETWORKS

We plot in Figs. 2,3, 4 three web-sites networks of the *memetracker* data-set. Reconstructions have been performed using information coming from only those cascades that satisfy the constraints described in the Results section and that contain the words "golf", "CIA" and "Syria" in the reference sentence.

ACKNOWLEDGMENTS

We warmly thank L. Dall'Asta for useful discussions, and Riccardo Refolo for providing us with Fig. 1. AB and APM acknowledge support by Fondazione CRT, project SIBYL under the initiative "La Ricerca dei Talenti", INFERNET, European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 734439, and PRIN project 2015592CTH_003 from the Italian ministry of university and research.

AUTHOR CONTRIBUTIONS

AB, AI and APM contributed equally to this work.

[1] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, and Riccardo Zecchina. Large deviations of cascade processes on graphs. *Physical Review E*, 87(6):062115, June 2013. doi:10.1103/PhysRevE.87.062115.

[2] F. Altarelli, A. Braunstein, L. Dall'Asta, and R. Zecchina. Optimizing spread dynamics on graphs by message passing. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09011, September 2013. ISSN 1742-5468. doi:10.1088/1742-5468/2013/09/P09011.

[3] Fabrizio Altarelli, Alfredo Braunstein, Luca Dall'Asta, Alessandro Ingrosso, and Riccardo Zecchina. The patient-zero problem with noisy observations. *Journal of Statistical Mechanics: Theory and Experiment*, 2014(10):P10016, 2014.

[4] Jonathan S Yedidia, William T Freeman, and Yair Weiss. Bethe free energy, kikuchi approximations, and belief propagation algorithms. *Advances in neural information processing systems*, 13, 2001.

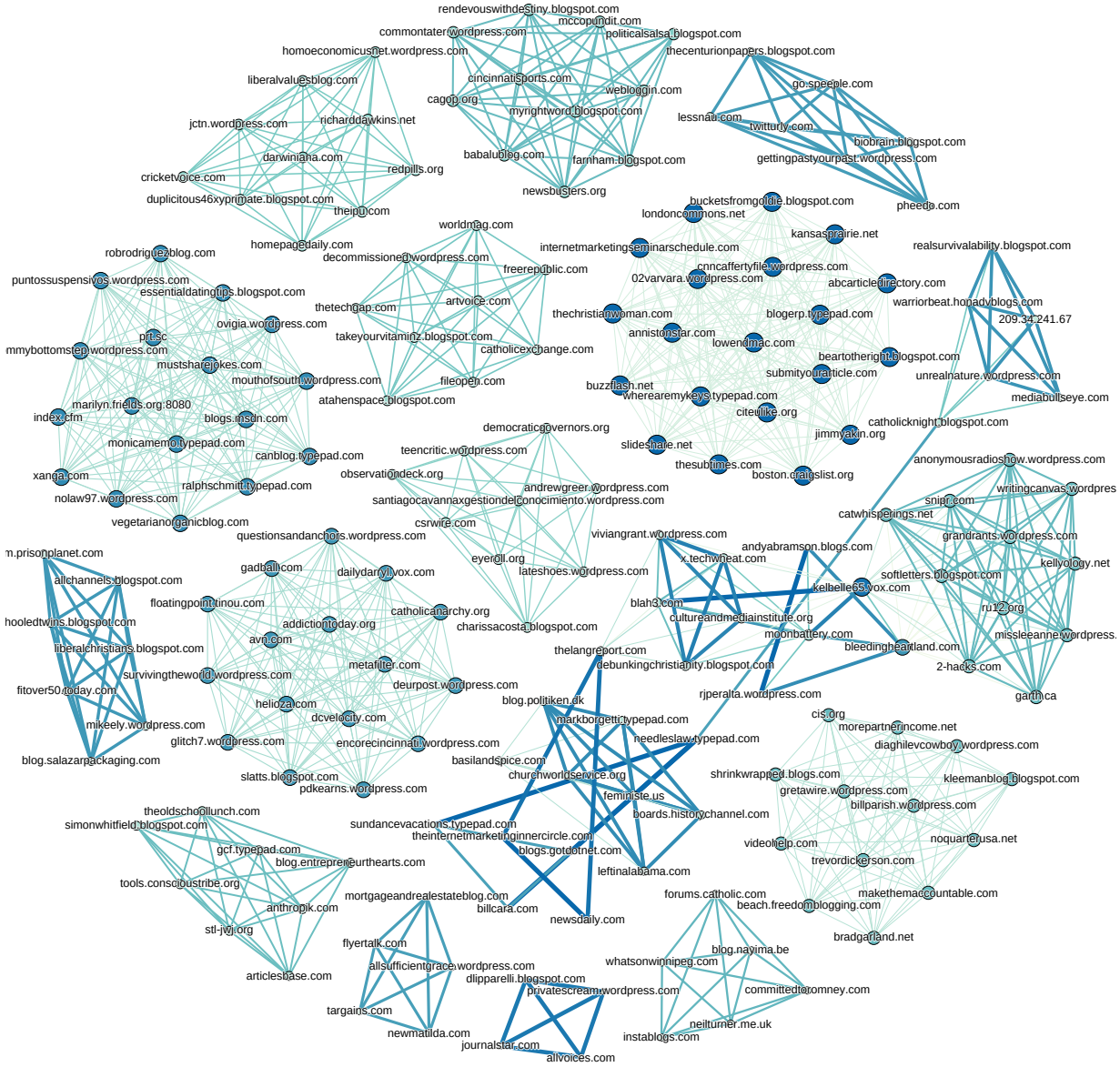

Figure 2. Web-pages networks publishing trend topics containing the word “golf”. Here $M = 24$, $|V| = 183$ and $|E| = 953$.

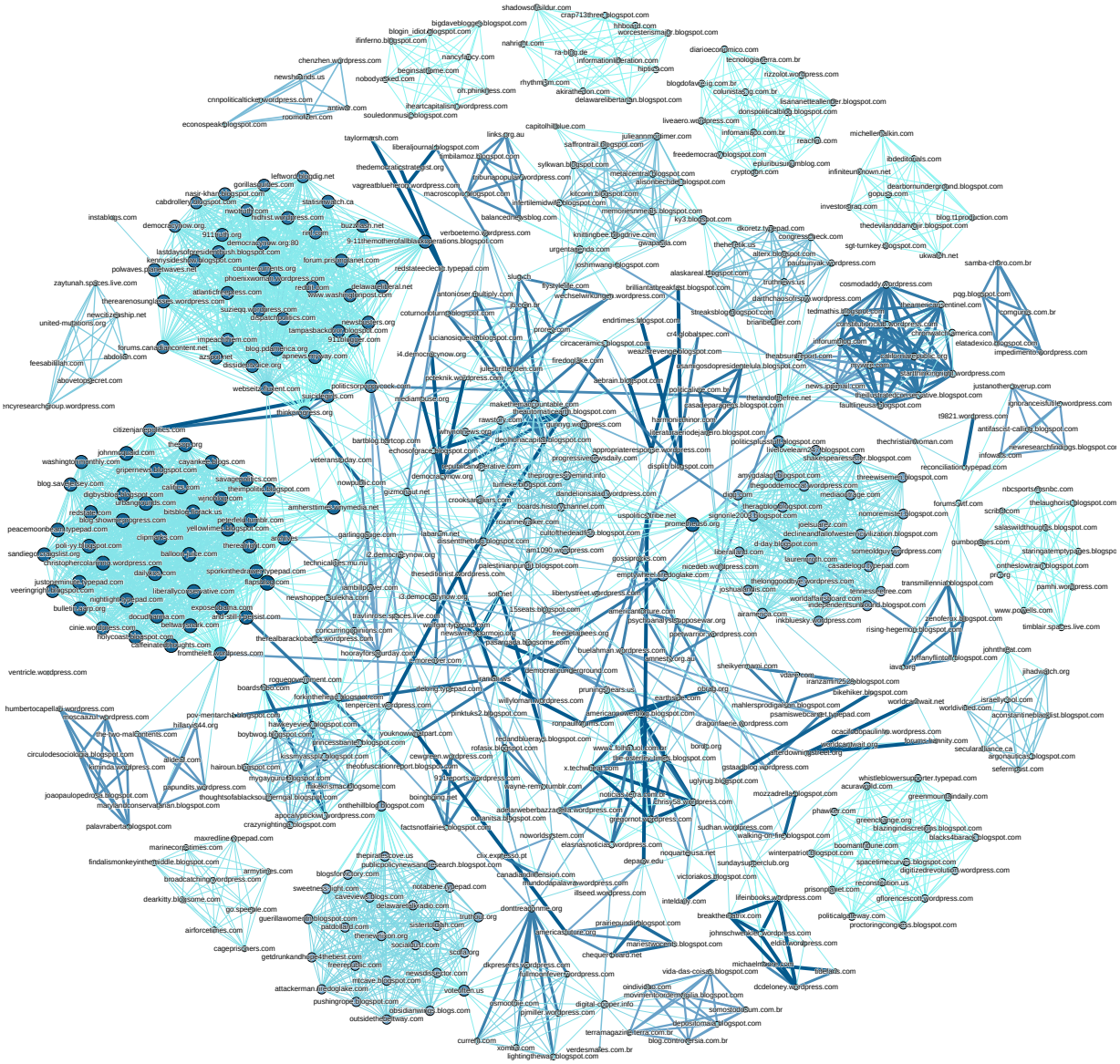

Figure 3. Web-pages networks publishing trend topics containing the word “CIA”. Here $M = 58$, $|V| = 491$ and $|E| = 3814$.

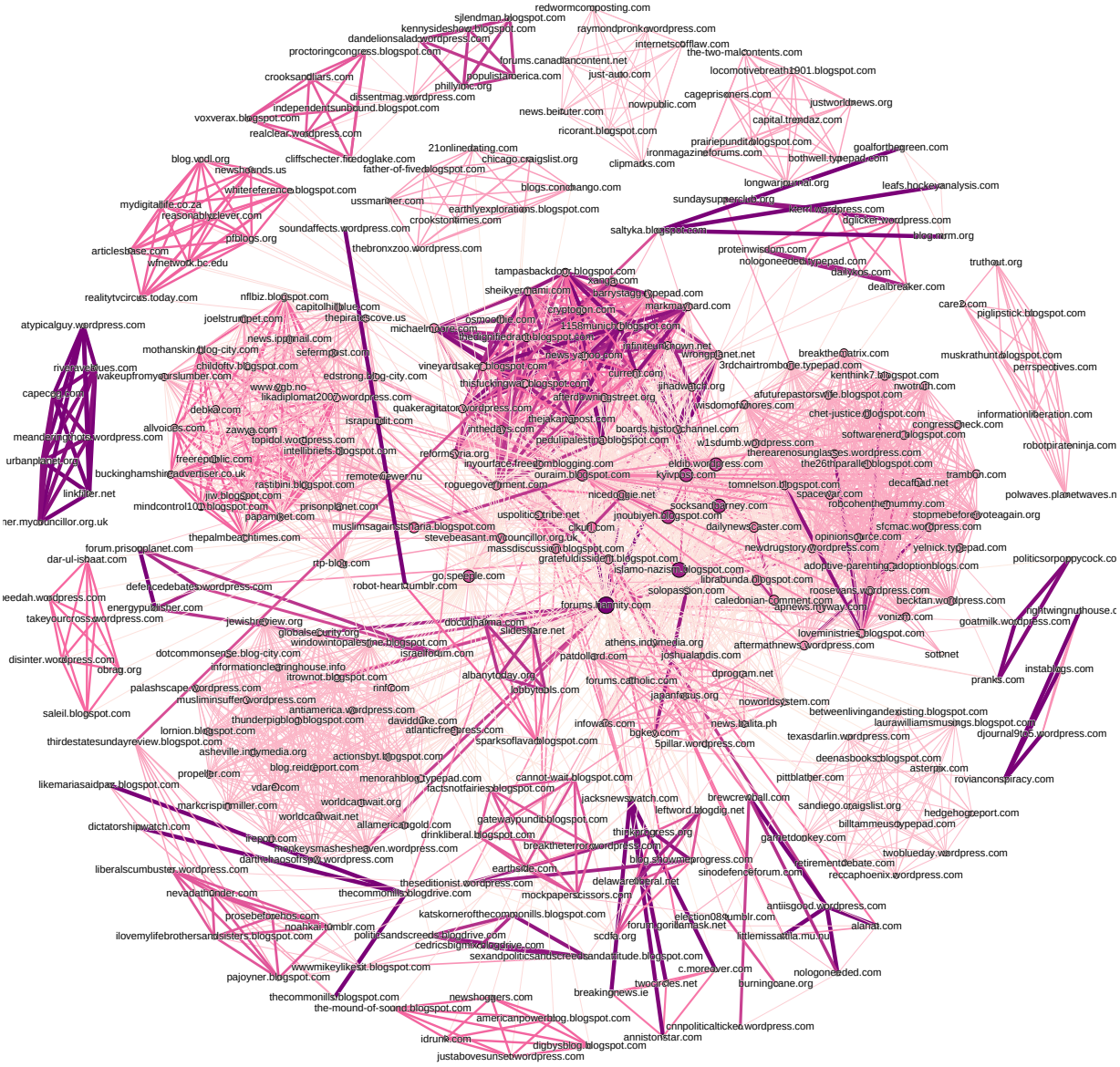

Figure 4. Web-pages networks publishing trend topics containing the word “Syria”. Here $M = 41$, $|V| = 302$ and $|E| = 3183$.