A Latent Parameter Node-Centric Model for Spatial Networks

Nicholas D. Larusso*, Brian E. Ruttenberg, Ambuj Singh

Department of Computer Science, University of California Santa Barbara, Santa Barbara, California, United States of America

Abstract

Spatial networks, in which nodes and edges are embedded in space, play a vital role in the study of complex systems. For example, many social networks attach geo-location information to each user, allowing the study of not only topological interactions between users, but spatial interactions as well. The defining property of spatial networks is that edge distances are associated with a cost, which may subtly influence the topology of the network. However, the cost function over distance is rarely known, thus developing a model of connections in spatial networks is a difficult task. In this paper, we introduce a novel model for capturing the interaction between spatial effects and network structure. Our approach represents a unique combination of ideas from latent variable statistical models and spatial network modeling. In contrast to previous work, we view the ability to form long/short-distance connections to be dependent on the *individual* nodes involved. For example, a node's specific surroundings (e.g. network structure and node density) may make it more likely to form a long distance link than other nodes with the same degree. To capture this information, we attach a latent variable to each node which represents a node's *spatial reach*. These variables are inferred from the network structure using a Markov Chain Monte Carlo algorithm. We experimentally evaluate our proposed model on 4 different types of real-world spatial networks (e.g. transportation, biological, infrastructure, and social). We apply our model to the task of link prediction and achieve up to a 35% improvement over previous approaches in terms of the area under the ROC curve. Additionally, we show that our model is particularly helpful for predicting links between nodes with low degrees. In these cases, we see much larger improvements over previous models.

Citation: Larusso ND, Ruttenberg BE, Singh A (2013) A Latent Parameter Node-Centric Model for Spatial Networks. PLoS ONE 8(9): e71293. doi:10.1371/journal.pone.0071293

Editor: Sergio Gómez, Universitat Rovira i Virgili, Spain

Received October 18, 2012; **Accepted** June 30, 2013; **Published** September 23, 2013

Funding: This research was sponsored in part by the U.S. National Science Foundation under grant IIS-0808772 and by the Army Research Laboratory under cooperative agreement W911NF-09-2-0053 (NS-CTA). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: nlarusso@cs.ucsb.edu

Introduction

Network analysis has been successfully applied to several scientific fields of study including sociology [1–3], information science [4,5], and ecology [6,7]. In many cases, the spatial configuration of nodes is paramount in analyzing a network as it plays a significant role in the formation and maintenance of links. Despite the important relationship between space and structure, many models and analyses are limited to only the network topology. Obviously such models fail to capture important *spatial* properties inherent in the data [8–10]. For example, in transportation networks, it is more economical to create short links between nodes [11,12]. Similarly, users in a social network are more likely to form links based on physically proximity because they have more interaction opportunities [3,13].

Although a plethora of spatial network models have been introduced in the literature (e.g. [3,14–18]), they assume that there is only one global link-cost function over the entire network, and it is a function only of distance. For instance, the exponential distance model [15,18] defines the probability of node i connecting to node j as $p(A_{ij}=1)=\frac{k_i k_j}{Z} exp(-d_{ij}/\hat{d})$, where the single parameter, $\hat{d}$, is set to the average pairwise distance between all nodes that share a link. Such models assume that the only node-specific influence on forming connections is the degree.

We test the fit of an exponential distance decay function on four real-world spatial networks: *C. elegans* neuron connections, social connections between users in Gowalla (a social photo sharing service), Internet server connections within California, and an airline transportation network for the United States (details provided in table 1). We show the distribution of the pairwise distances of connected nodes in figure 1, as well as a maximum likelihood fit to an exponential distribution. Although we see that only the Gowalla network potentially fits well to an exponential distribution, we perform a Kolmogorov-Smirnov (KS) test on each network to quantitatively test the fit. In fact, all of the networks reject the null hypothesis (that the data come from the same distribution) with p-values $4.6e^{-152}$ (*C. elegans*), $2.2e^{-6}$ (Gowalla), $1.7e^{-55}$ (CA Internet), and $5.2e^{-29}$ (US Airline).

Additionally, the *C. elegans* and CA Internet networks contain a small second mode in the tail of the distribution, caused by areas of heavy spatial clustering of the nodes. This tight interaction between the spatial distribution of nodes and the likelihood of observing long-distance connections makes it difficult to describe the distance with a single function over the entire network.

In this paper, we investigate the variable effects of distance on *individual* nodes and how this influences network topology. To

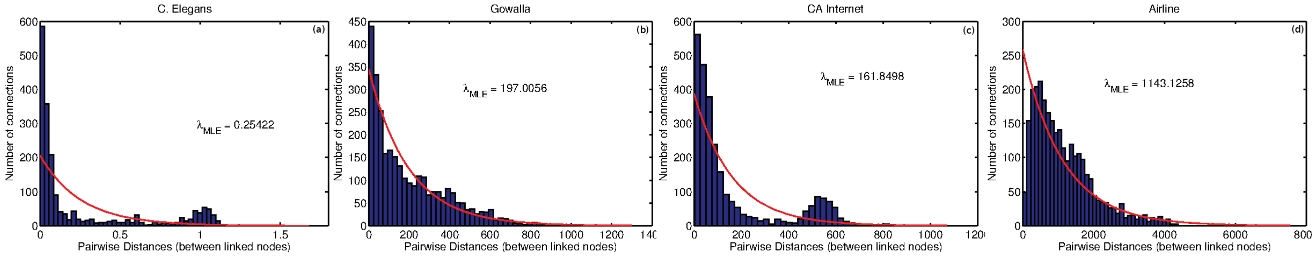

Figure 1. Distribution of the pairwise distances between linked nodes along with a maximum likelihood fit to an exponential distribution.
doi:10.1371/journal.pone.0071293.g001

model these effects we combine ideas from previous spatial network models [15,17,18] with latent parameter models [19,20]. We capture the effect of distance on each individual node's ability to form links of various lengths by attaching a latent radius parameter to each node. Furthermore, we extend this idea by adding a second node-specific latent variable which captures space-independent community structure. Our experiments show that our model achieves up to 35% improvements over other methods in the task of link prediction (in terms of area under the ROC curve). Moreover, we see the most significant improvements (up to 80%) when predicting links between nodes with low degrees, where many link prediction techniques fail.

Related Work

The development of mathematical models of network structure has played an important role in advancing the area of network science [4,5,21–25]. In this section we review the relevant research work in the areas of spatial network models and analysis and statistical network models.

Spatial Networks

The existing work on modeling spatial networks can be split into three general types of models: (i) Waxman models, (ii) geometric models, and (iii) preferential attachment and scale free spatial models. Perhaps the earliest model to incorporate the pairwise distance between nodes into the probability of a link was the Waxman model [26]. Specifically, the authors proposed that the probability of a link is proportional to $Be^{-d_{ij}/L}$, for some constant B and scaling coefficient L. The Waxman model can be construed as the spatial equivalent of the Erdos–Renyi random graph model (ER) [22] since as $L \rightarrow \infty$, the model converges to the ER attachment model. While this spatial model has been shown to replicate some real world networks (e.g. [27]), it fails to capture the preferential attachment that has been observed in a variety of networks [4,5,11,12].

Table 1. Properties of the real-world spatial network datasets we examine in this paper [59].

Name	Type	Nodes	Edges	Area	Index of dispersion
C. elegans	Biological	277	1,918	0.012 μm^2	7.163
Gowalla	Social	600	340	776,000 km^2	23.098
Internet	Infrastructure	501	2,661	809,000 km^2	11.317
US Airline				Transportation	476
2,773				16,140,695 km^2	1.564

The last column refers to the *index of dispersion*, a measure of complete spatial randomness (CSR) of the nodes [65]. Values close to 1 indicate that the nodes are likely to be distributed uniformly over the space, whereas values greater than 1 result from too little dispersion (e.g. nodes tend to cluster in space).
doi:10.1371/journal.pone.0071293.t001

The class of *geometric* models, describe the probability of a link forming between two nodes as a function of distance which approaches one as the distance between two nodes decreases. Typically the probability of attachment is formulated as a logistic, $\frac{1}{1+e^{-A(d_{ij}+B)}}$, where A is a scale parameter controlling the slope of the logistic and B controls the shift of the function. Pure geometric networks, where an edge between two nodes exists if the distance is less a certain threshold, can be considered a special case of a logistic function with $A \rightarrow \infty$. Many works have studied the theoretical network statistics of these thresholded graphs under the assumption of uniform spatial distribution [28,29]. Additionally, Wong et. al. [3] propose a similar logistic spatial model for social networks that replicates several statistic of real world networks.

Several existing network models have been adapted to incorporate spatial information as well. Typically, the probability of attachment in these networks is proportional to $k_i e^{-d_{ij}/L}$ or $k_i d_{ij}^A$, such that one is able to generate random networks with a given expected degree distribution, where the probability of any two nodes forming a link decays exponentially or as a power law with distance [14,18]. Properties of these networks have been well studied [30–32], particularly that as L and A vary, the structure of the spatial networks can change from scale–free networks with little clustering to large networks with intense clustering [30]. While these models are adept at modeling the evolution of complex spatial networks such as the Internet [33], they still assume a homogeneous spatial effect throughout the network.

In addition to modeling, several authors have studied the structural properties of spatial networks and understand the role that space plays in the network topology. Specifically, there has been a large amount of work merging traditional network models with spatial models, and determining how these network models change under spatial constraints [8–10,14,30,34,35]. For instance, in [10], the authors discuss how scale–free networks can be analyzed in a geometric space. The resulting models can be applied to several types of data to analyze the structural properties and provide insight into the link creation process. Such analyses are especially important in understanding biological networks [27,36].

The distribution of nodes in space also affects the types of connections, and therefore the global structural properties of a spatial network. Bullock et al. [37] discuss several properties of spatial networks and how the spatial distribution of the nodes effect these properties. For instance, when nodes are distributed uniformly in a given space, there is a sharp phase transition in the size of the largest component of the network, whereas nodes distributed in an inhomogeneous manner, exhibit a smooth transition in the number of connected components and their sizes. Additionally, Voges et al. [38] study the network properties (e.g. degree correlation, shortest path length, cluster coefficient, and spatial concentration) of networks embedded into a lattice. The authors experimented by adding some jitter to the node positions and studying the resulting of network statistics. They found that these properties are very sensitive to the randomness of the node locations. This further corroborates the importance of including the spatial properties of networks when studying their structural properties.

Beyond analyzing the structure of spatial networks, recent approaches to community detection in spatial networks propose new null network models, based on gravity models [39], which are implemented within the modularity framework [40]. The idea is to incorporate the pairwise distance between nodes into the expectation of whether or not a link exists between them, thus more accurately representing the spatial network structure [15,17]. In Cerina et al. [15], the authors propose a model in which the probability of a link forming between two nodes declines exponentially as the distance between them increases. In Expert et al. [17], the authors build an empirical distribution of the probability of connection conditioned on the distance from the observed network and use that to weight the connection probability. In both cases, the authors assume that the effect of distance remains constant throughout the entire network. Both of these models have shown to improve community findings in spatial networks over the originally proposed null model (i.e. the configuration model: $\frac{k_i k_j}{2\sum_t k_t}$).

In addition to descriptive modeling, Lennartsson et al. [41] introduce *SpecNet*, a general spatial network model that is capable of *generating* networks with a full range of values for clustering coefficient, degree assortativity [42], and fragmentation index. Whereas previous models were only able to create networks with a very limited range of possible statistics, *SpecNet* is able to produce networks that can nearly cover the range of possible theoretical values for such measures. Such generative models provide a more concrete link between the various components of the network and how these relate to the structural properties.

Latent Parameter Network Models

Hoff et al. [19] introduce a latent space approach for modeling social networks. The authors construct a model in which the objective is to infer node positions in a *latent social space* such that links are more likely between nodes that are close together in this latent space. In fact, given each nodes' location in this latent social space, all of the network links are conditionally independent. This model is able to effectively represent a large number of social networks due to its ability to capture homophily. That is, nodes close together in latent space typically have similar distances to other nodes as well. Others have introduced interesting theoretical properties of this model as well as offered their own extensions [43–45]

Additionally, Hoff et al. [20,46,47] have further developed more general latent factor models which have been shown to generalize [19]. In [20,47], the basic idea is to model network connections as $y_{i,j} \propto \beta X + uDu^T$, such that each link is a function of a set of covariates as well as a low rank approximation of node-wise random effects. The authors show that this model weakly generalizes the latent space and class models previously proposed, and provides high quality predictions for a wide variety of networks (e.g. social networks, word relationship networks, and protein interactions). In contrast, our objective in this work is to separate the set of dependent variables such that we isolate the *spatial term* from the others. As our hypothesis is that spatial effects vary over the network, we want to study the effect on each node in the original space.

Related are the class of exponential random graph models or p^* graphs [48,49]. The p^* model is a probability distribution over networks that takes the form: $p(Y|\theta) = \frac{exp(\theta s(y))}{c(\theta)}$, where Y is the adjacency matrix of the network, $s(y)$, is a vector of sufficient statistics over the network, θ is a vector of parameters, and $c(\theta)$ is a normalization term which depends on the parameter vector. Because the statistic vectors, $s(y)$, are collected over the entire network, p^*, models are often used to capture complex interactions in social networks. However, this flexibility comes at a severe cost in that the normalization constant is typically intractable, containing a sum over an exponential number of parameter configurations, greatly complicating parameter estimation and inference algorithms for these models [48].

Lastly, block models are another form of latent variable models, often used for community detection, in which each node is associated with a latent group parameter such that nodes are more likely to form connections within a group than between groups [50–52]. These models assume nodes fall into equivalence classes such that the probability of a pair of nodes connecting is conditionally independent given the latent group identifiers of nodes. The inferential problem is then to compute the latent class identifier for each node, given the network structure. For a more

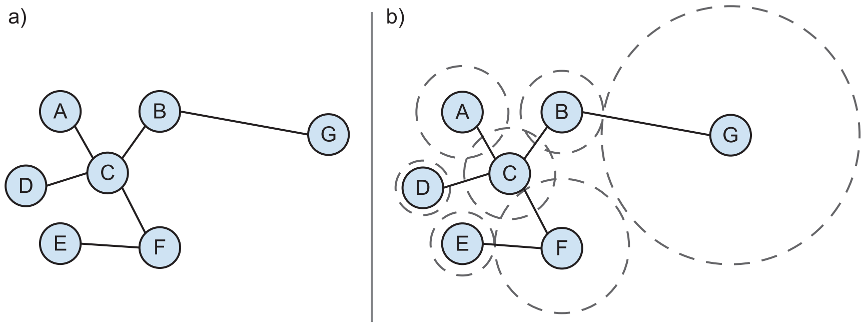

Figure 2. Illustration of how the radii summarize the local network and spatial structure for each node.
doi:10.1371/journal.pone.0071293.g002

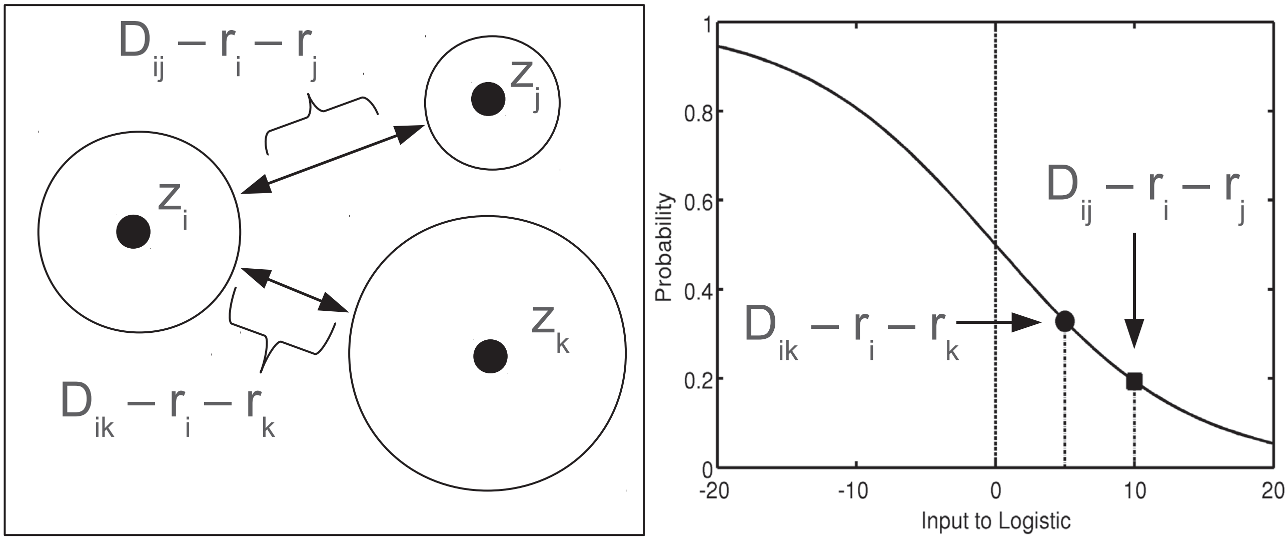

Figure 3. Illustration of how the radii from different nodes interact with each other and the pairwise distance to determine the existence of an edge.
doi:10.1371/journal.pone.0071293.g003

comprehensive survey of the work in this area, we refer the reader to [1].

Additionally, there has been research in embedding complex networks into an underlying, often hidden, metric or geometric space [53–57]. For example, Papadopoulos et al. [56] introduce a geometric interpretation of the preferential attachment model. The authors embed a network into a geometric space, and introduce a model that combines node similarity, defined by closeness in space, and popularity in order to compute linkage probabilities. As a new node is added to the network, it selects a subset of existing nodes to which it will connect proportional to the hyperbolic distance between the nodes [58]. This ensures that nodes connect to others which are not only popular, as in the case of preferential attachment, but also similar, as defined by their pairwise distance. As with other latent variable models, these approaches aim to embed an observed network into a hidden space such that pairwise distances capture node similarity whereas our objective in this work is to understand the relationship between node distances in their true Eulidean space with the resulting network structure.

Methods

In this section we introduce a novel probabilistic model for analyzing spatial networks in which spatial effects are captured at the level of individual nodes. To capture the variable effects of space throughout the network, we introduce a latent, positive real-valued, parameter referred to as the radius at each node. We introduce two models which incorporate this idea, *Radius* and *Radius+Comms*. The first model, *Radius*, only models the node-specific spatial effects and node popularity. The second model, *Radius+Comms*, adds a component to capture community structure within the network which cannot be explained by factors incorporated in the *Radius* model.

Figure 2 illustrates how the radius summarizes the local structure and node density in the *Radius* model. First, we notice

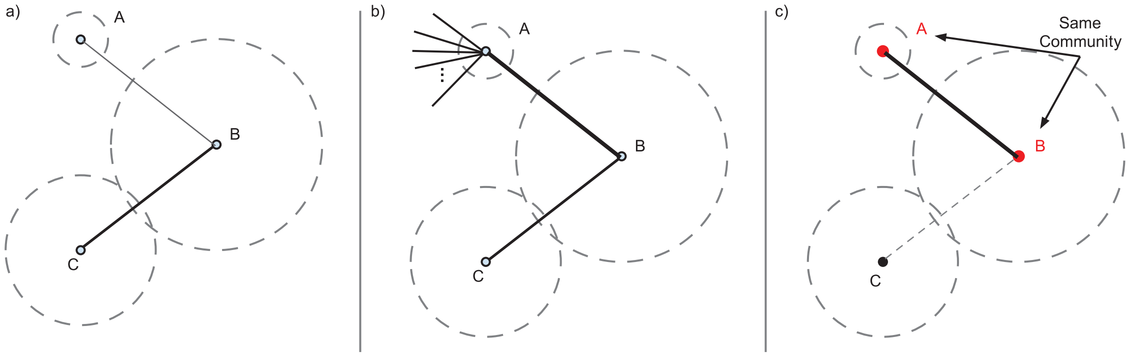

Figure 4. The different mechanisms that may influence the probability of a connection between two nodes. In each of the instances, the distance from node *A* to *B* and from node *C* to *B* are equal. In figure (a) the link probabilities are determined by the combined radii of the nodes. It is much more likely that nodes *B* and *C* will form a link due to their radii. In figure (b), the probably of a link between nodes *A* and *B* increases because node *A* is a hub (i.e. high node degree), even though it still has a small spatial reach. In figure (c), nodes *A* and *B* have a high probability of forming a link because they are both in the same community. In contrast the probability of a connection between *B* and *C* is reduced because they are in different communities.
doi:10.1371/journal.pone.0071293.g004

the tight cluster of nodes, $A-F$, all have fairly small radii. This is because these nodes only share links with a small subset of other nodes within the cluster, thus there is a strong penalty for growing the radius of any node too large. Second, we see that although node G has only one connection, it exhibits a large radius. This is due to the observed connection with node B and the fact that B's radius is restricted by node A. Given the low degree for each of these nodes, node G's radius must overcome the distance gap to node B in order to explain this connection. Lastly, we see that although there is a connection between nodes C and D, there is a significant gap between the nodes' radii. This is due to the fact that node C has a relatively high degree, thus this connection can be partly explained away, reducing the constraints on the radii. Note that we omit uncertainty in the size of the radii here for clarity, though we would expect that the radii of nodes in dense regions to have lower variability and nodes in regions of low density to have higher variability.

Throughout this work, we assume that we are given as input a spatial network. A network is represented by the adjacency matrix, A, where $A_{ij}=1$ if there is a link between nodes i and j. The pairwise distances between nodes is given by the matrix, D, such that D_{ij} is the Euclidean distance between nodes z_i and z_j. The degree, or importance, vector for the nodes, K, is considered in our model as a constant (since it is always conditioned upon) and provides a measure of how likely it is that this node will take part in newly formed links given its current popularity. That is, since our objective is to predict *new* potential links given a partially observed network, we do not enforce $k_i=\sum_{j\neq i} A_{ij}$ on the inferred values of A_{ij}.

Basic Spatial Model: *Radius*

The *Radius* model is based on the idea that space may influence each node differently. The model consists of two terms, (i) a spatial term which favors creating a link between nodes when their radius-corrected pairwise distance is small and (ii) an observed *popularity* term which, all things equal, favors linking nodes that already have many connections. We combine both of these terms within the logistic function since the output is interpreted as the probability of an edge existing between two nodes (i.e. a binary outcome, either an edge exists: $A_{ij}=1$, or it does not: $A_{ij}=0$). The probability of a link is defined in Eq. 1.

$$
\begin{aligned}
&p(A_{ij}|r_i,r_j,D_{ij},K,\alpha,\gamma)\\
&=\frac{1}{1+exp\left(-\frac{1}{\alpha}(r_i+r_j-D_{ij})+\frac{1}{\gamma}\left(\frac{k_ik_j}{\sum_z k_z}-M\right)\right)}
\end{aligned}
\tag{1}
$$

The first term, $\frac{1}{\alpha}(r_i+r_j-D_{ij})$, describes the propensity of a pair of nodes to form a link given their (latent) radius parameters and the distance separating them. Although it is more costly to form long distance links in general, the radii can reduce or even completely overcome this cost. The scale parameter, α, controls the strength of the distance term on the overall link probability. This parameter also allows the model to automatically adapt to networks at different scales.

Figure 3 illustrates the role of the radii in forming a link between two nodes separated by distance, D_{ij}. Although nodes may be separated by a large distance, if the combined radii can make up for this distance, or at least reduce it, a link between these nodes becomes more likely. That is, we assume a simple linear relationship between radii and pairwise distance: $D_{ij}-r_i+r_j$. Since we would like to predict the output of $0/1$, depending on whether an edge exists or not, we place this term into a logistic function.

The second term describes the propensity of nodes to form links with popular nodes (i.e. nodes with a large degree). This is the standard term considered in configuration-based models [40] for adding network edges in a manner that is proportional to the current node popularity (i.e. the rich get richer). The constant M is defined as the midpoint between the average combined degrees of the set of pairwise nodes that share an edge and those that do not. That is, if $k_xk_y<M<k_ik_j$, then, given no other information, we would expect $p(A_{ij})>p(A_{xy})$ simply due to the fact that there are more possibilities for these two to share an edge. More formally, we compute the value of M from the network as show in Eq. 2.

$$
\begin{aligned}
S&=\left\{\frac{k_ik_j}{\sum_z k_z}|A_{ij}=1\right\}\\
\hat{S}&=\left\{\frac{k_ik_j}{\sum_z k_z}|A_{ij}=0\right\}\\
M&=\frac{1}{2}\left(1/|S|\sum_i S_i+1/|\hat{S}|\sum_i \hat{S}_i\right)
\end{aligned}
\tag{2}
$$

Including this constant allows this term, $\frac{k_ik_j}{\sum_z k_z}-M$, to take on both positive and negative values. Because this term is placed into a logistic function, this allows us to both increase and decrease the probability of a link based on the combined node degrees. Thus providing a mechanism to explain the existence of links due to the popularity of the nodes or the absence of a link due to the fact that the nodes have few observed connections.

The parameter, γ, is again a scaling parameter which controls the total influence of this term on the resulting link. The two scaling parameters offer a large degree of flexibility to the model since it is able to automatically adapt to networks with both very strong and very weak spatial effects.

The posterior distribution for our model is given in Eq. 3. Our objective is to infer values of the hidden variables, α, γ, and R (the vector of radii), given the observed network structure, A, node degrees, K, and pairwise distances, D. We use truncated Gaussian distributions, denoted $\mathcal{N}_{>0}()$, for priors over all of the latent variables in our model (since all of the variables are restricted to be positive). We discuss the inference computation more in section.

$$
\begin{aligned}
p(R,\alpha,\gamma|A,K,D)\propto p(A|R,D,K,\alpha,\gamma)p(R)p(\alpha)p(\gamma)&\\
=p(\alpha)p(\gamma)\prod_{i>j}^{n}p(A_{ij}|r_i,r_j,D_{ij},k_i,k_j,\alpha,\gamma)p(r_i)p(r_j)&\\
=\mathcal{N}_{>0}(\alpha;\mu_\alpha,\sigma_\alpha)\,\mathcal{N}_{>0}(\gamma;\mu_\gamma,\sigma_\gamma)&\\
\times\prod_{i>j}^{n}logistic\left(\frac{1}{\alpha}((r_i+r_j)-D_{ij})+\frac{1}{\gamma}\left(\frac{k_ik_j}{\sum_z k_z}-M\right)\right)&\\
\times\mathcal{N}_{>0}(r_i;\mu_r,\sigma_r)\,\mathcal{N}_{>0}(r_j;\mu_r,\sigma_r)&
\end{aligned}
\tag{3}
$$

Community Model: *Radius+Comms*

Although nodes that are physically close together are more likely to form a link than nodes that are further apart, space is not the only factor in deciding which nodes should be connected. Previous literature [1,14] often identify three main explanations of links: (i) close spatial proximity, (ii) node popularity, and (iii) community structure within the network. These factors are illustrated in figure 4.

Table 2. Metropolis within Gibbs sampling routine for Bayesian inference of our spatial network model.

```
$\alpha,\gamma,r_i \forall i$// Randomly initialize random variables:
for $s=1\rightarrow T$ do
// propose new values for global variables
$\hat{\gamma}\sim\mathcal{N}(\gamma_{s-1},\sigma_\gamma)\hat{\alpha}\sim\mathcal{N}(\alpha_{s-1},\sigma_\alpha)$,
// compute acceptance ratio
$\text{acceptpatio}=(logP(A|R,D,\hat{\alpha},\hat{\gamma})+logP(\hat{\alpha})+logP(\hat{\gamma}))-(logP(A|R,D,\alpha^{s-1},\gamma^{s-1})$
$+logP(\alpha^{s-1})+logP(\gamma^{s-1}))$
$u\sim$ unif(0, 1)
if $log(u)<$ acceptRatio then
$\alpha^s=\hat{\alpha}$, $\gamma^s=\hat{\gamma}$ // accept samples
else
$\alpha^s=\alpha^{s-1}$, $\gamma^s=\gamma^{s-1}$ // reject samples
end if
// propose new values for node variables
for $j=1\rightarrow n$ do
$\hat{r}_i\sim\mathcal{N}(r_i^{s-1},\sigma_r)$
$=(logP(A_{i-}|R_{-i},\hat{r}_i,D,\alpha,\gamma)+logP(\hat{r}_i))-(logP(A_{i-}|R_{-i},r_i^{s-1},D,\alpha,\gamma)+logP(r_i^{s-1}))$acceptRatio
$u\sim$ unif(0, 1)
if $log(u)<$ acceptRatio then
$r_i^s=\hat{r}_i$ // accept sample
else
$r_i^s=r_i^{s-1}$ // reject sample
  end if
  end for
end for
```

doi:10.1371/journal.pone.0071293.t002

With the basic model in place, we develop an extension, *Radius+Comms*, which allows us to simultaneously infer any space-independent community structure within the network as well. To describe the community structure, we attach a discrete latent parameter to each node which identifies the node's group label, $c_i \in \{0,\ldots,K\}$. Nodes within the same community should have more links to other nodes within their community and fewer links to nodes in other communities. We model this by adding a (latent)

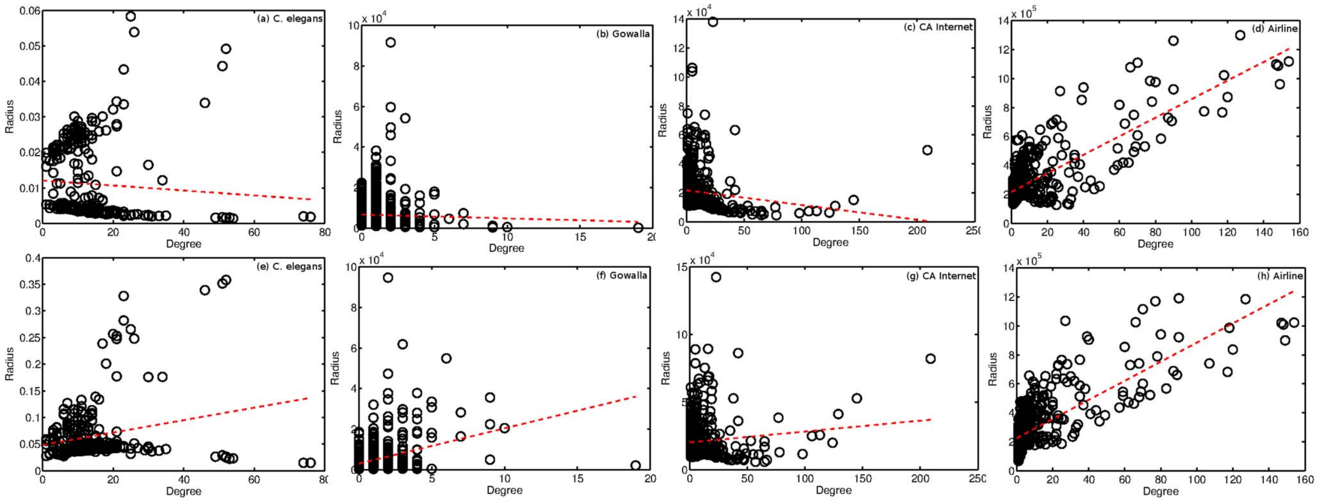

Figure 5. Degree versus mean posterior radius for each network. The dotted line in each figure is the ordinary least squares regression fit to this data, where degree is the covariate and radius is the response (i.e. $radius=m\ degree+b$). The Pearson correlation between mean posterior radius and degree for the *Radius* (*Radius+Comms*) model for each network is (a) $-0.07(0.23)$, (b) $-0.03(0.32)$, (c) $-0.14(0.11)$, and (d) $0.78(0.77)$.
doi:10.1371/journal.pone.0071293.g005

random variable within the logistic function. This way the community effects do not completely override spatial behavior of nodes, rather they can strengthen or dampen the effects of distance on a particular connection to make it a more probable outcome.

Unlike most community detection methods, we offer a *don't care* community ($c_i = 0$) which allows the formation of links between nodes to follow only the previously described model. That is, for nodes placed into the *don't care* community, the probability of a link involving this node remains unchanged, even if the link connects to a node in another community. This formulation ensures that our model will only capture salient network structure which cannot otherwise be explained by other factors. The new community term, $\beta(c_i, c_j)$, is given in Eq. 4.

$$\beta(c_i, c_j) = \begin{cases} 0 & c_i = 0 \text{ or } c_j = 0 \\ \phi & c_i = c_j \\ -\phi & c_i \neq c_j \end{cases} \tag{4}$$

If nodes belong to the same community, we increase the probability of a connection by adding ϕ to the other terms within the logistic function. Where ϕ is a positive, real-valued random variable to be inferred from the observed data. Combining this with our previous model, the updated posterior distribution is given in Eq. 5.

$$\begin{aligned} p(R,C,\alpha,\gamma,\varphi|A,K,D) \propto & \mathcal{N}_{>0}(\alpha;\mu_\alpha,\sigma_\alpha)\, \mathcal{N}_{>0}(\gamma;\mu_\gamma,\sigma_\gamma)\, \mathcal{N}_{>0}(\varphi;\mu_\varphi,\sigma_\varphi) \\ & \times \Pi_{i>j}^{n} logistic \begin{pmatrix} \frac{1}{\alpha}((r_i + r_j) - D_{ij}) \\ + \beta(c_i, c_j) \\ + \gamma\left(\frac{k_i k_j}{\sum_z k_z} - M\right) \end{pmatrix} \\ & \times \mathcal{N}_{>0}(r_i;\mu_r,\sigma_r)\, \mathcal{N}_{>0}(r_j;\mu_r,\sigma_r) \\ & \times \text{multinomial}(c_i;\theta_C)\, \text{multinomial}(c_j;\theta_C) \end{aligned} \tag{5}$$

The new random vector, C, encodes the community IDs for each node and if $c_i = 0$, then this node is assigned to the *don't care* community. The interaction between nodes within the same community and across communities is modified by the function $\beta(c_i, c_j)$ which is defined in Eq. 4. This adds one extra weighting (positive, real-valued) variable, ϕ. If $c_i = c_j$, then a large value of ϕ will increase the probability of a link between the two nodes, whereas if $c_i \neq c_j$, then $-\phi$ will decrease the probability of a link. Note that this defines a symmetric relationship; within-group connections are strengthened by the same amount that between-group connections are penalized.

The number of clusters, K, should be set sufficiently large to accommodate any structure that may exist. Because we include a *don't care* community, the specific setting of K is not critical since, if there is insufficient evidence of clustering, nodes may simply be assigned $c_i = 0$. However, as K increases, the rate of convergence of our inference routine may slow, since it much search a larger discrete space. In our experiments, we set K to 10% of the number of nodes in the network. We have found that this provides a nice trade-off between flexibility and efficiency as confirmed by our analysis of the MCMC trace plots. In fact, many of the networks we have tested identify fewer communities, and only the *C. elegans* network places every node into a community.

Inference

To compute with our model, we employ a standard Markov Chain Monte Carlo (MCMC) algorithm for approximate inference. We chose to apply Bayesian inference rather than maximum likelihood or stochastic search optimization to ensure that all of the uncertainty was appropriately propagated throughout the model. Just as it is unlikely that there exists a single global function over distance which can accurately capture the effects over the whole network, we do not expect the inferred radius values to be exact measures of the nodes' spatial reach.

The sampling procedure iterates between proposing new global parameter values (i.e. scaling parameters) with new radius values. Algorithm Table 2 outlines the full MCMC algorithm for the *Radius* model. Inference on *Radius+Comms* is a straightforward extension of this algorithm where we also infer the value of ϕ, the global community penalty and reward as well as the c_i's, the group ID's for each node.

We use the notation $logP$ to refer to the log of the probability density function. The vector, R, is the set of all radii, whereas R_{-i} is all of the radiis except for r_i. We use truncated Gaussians for all of the prior distributions since all of the parameters are restricted to positive values. Additionally, we set the parameters for the prior distributions to be rather uninformative, though specific to each network due to the differences in distance scales across our datasets. Lastly, we have experimented with different block-updating schemes, however, the one presented here, in which we first update the global scaling parameters, then each of the node parameters provided relatively fast convergence and good mixing for all of the networks (more discussion on this in section.

Results

We experimentally evaluate our proposed model by applying it to the task of link prediction on four different real-world spatial networks (described in table 1) [59]. Furthermore, we offer additional analysis of the model parameters and present interesting interpretations by utilizing additional information about the network nodes.

Analysis of Inferred Radii

We first investigate the inferred radii in more detail. Our claim was that the radius was meant to capture a node's spatial reach. While this is related to the degree of a node, we show that the radius will contain additional, unique information about a node's propensity to take part in long (short) distance connections. To test this, we plot the mean posterior radius for each node against its degree and test the amount of correlation in these values. We do this for both models and compare our results, shown in figure 5.

From figure 5, we make three interesting observations. First, there is a large variance in the inferred radii values corroborating our claim that distances effect individuals in a different manner. For example, in the *C. elegans* network, we see clusters around different radii for nodes with similar degrees. This likely corresponds to the spatial clustering of neurons in both the head and the tail of the worm. Neurons in the head require a much smaller spatial reach since they have many potential connections within a short distance. Similarly, neurons in the tail also cluster spatially, however, to a lesser degree, thus requiring a slightly larger radius. We see a similar pattern in each of the networks, though to a lesser degree since connections in these networks are much more localized than in *C. elegans.*

Second, there is little correlation between node degree and mean posterior radius. This indicates that the inferred radius values are capturing the spatial tendencies of each node, rather

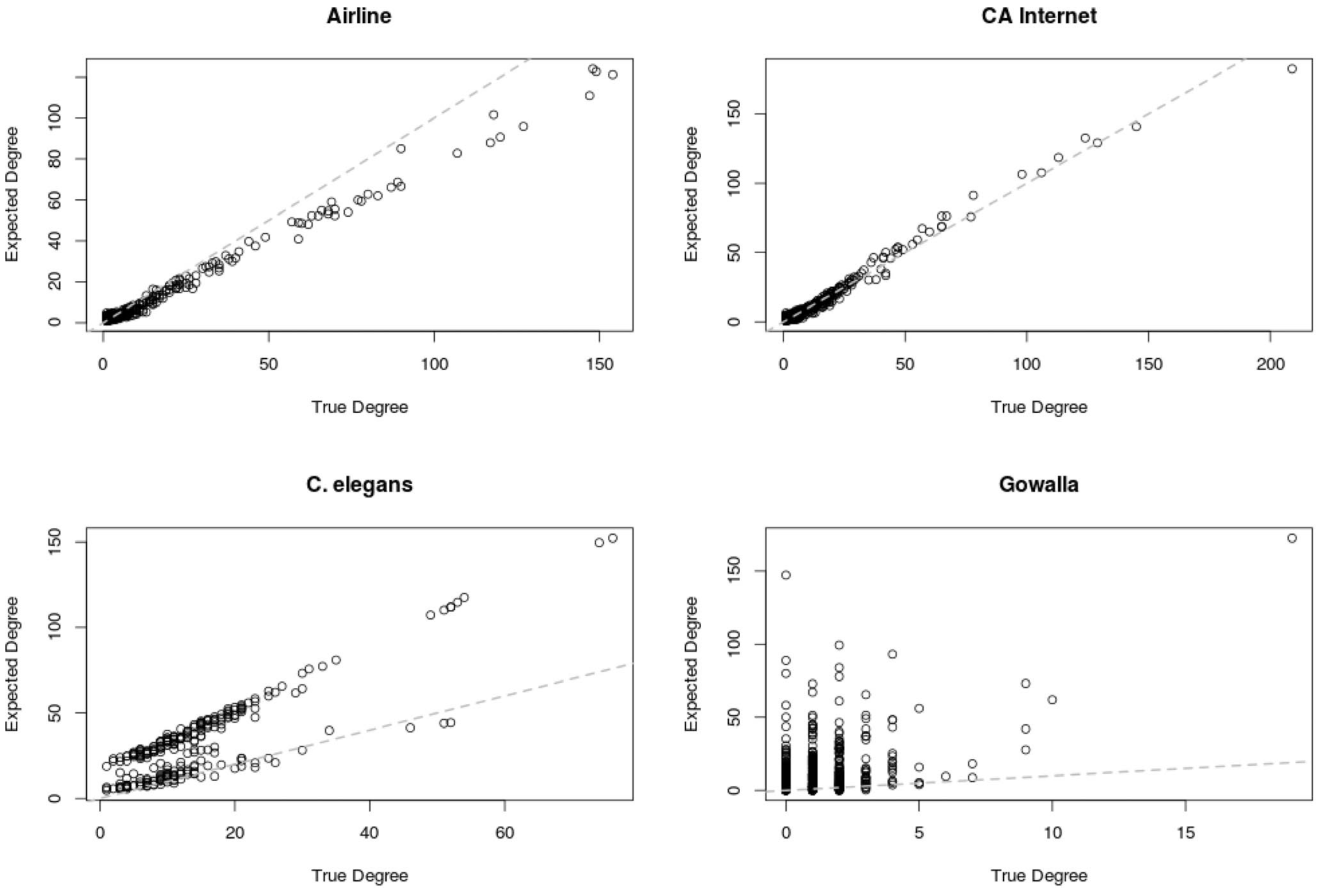

Figure 6. Node degrees from the original networks ploted against expected node degrees from the maximum a-posterior model parameters.
doi:10.1371/journal.pone.0071293.g006

than simply re-capturing a measure of node popularity. In fact, only the Airline network shows any significant correlation between these two values. We also notice that this is the only network for which the nodes are distributed nearly uniformly at random (see *index of dispersion* in table 1). When nodes are uniformly distributed, there will be little difference in any node's spatial reach since all nodes must extend approximately the same distance in order to reach another node. Thus nodes which take part in more connections will tend to extend further.

Third, the distribution of radii is different for the two models with no clear trend across all networks. The additional modeling power in *Radius+Comms* is used primarily to explain away the presence of abnormally long distance connections as well as the absence of closely co-located nodes of medium to high degree. In the first case, the radius for each of the nodes involved may be reduced since the abnormally long link is explained by an additional factor. In contrast, in the second case, the radii may grow larger, since the penalty of the two nodes belonging to different communities sufficiently explains why they do not connect. Depending on the particular network, we will likely see a mix of these two cases, thus causing some radii to grow and others to shrink accordingly.

Lastly, in figure 6, we compare the node degrees from the original networks to the expected node degrees using the maximum a-posterior (MAP) parameters. Overall, we see a strong correlation between the true and expected degrees, indicating that the model is congruent with the observed networks. In the Gowalla network, however, although a strong correlation does

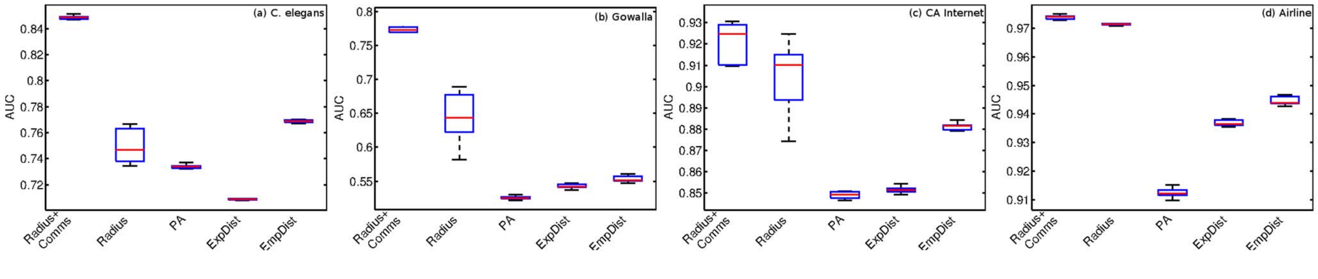

Figure 7. Link prediction AUC over 10-fold cross validation.
doi:10.1371/journal.pone.0071293.g007

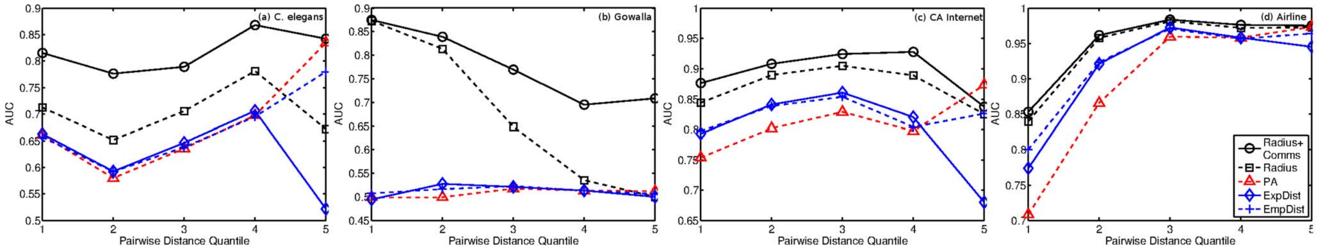

Figure 8. AUC measured over separate quantiles of the test data, split by the pairwise distance between the nodes for which a link is being predicted. The quantiles are shown on the x-axis, where 1 contains all node-pairs that are close together, and 5 contains those that are separated by the greatest distances.
doi:10.1371/journal.pone.0071293.g008

exist, $\rho = 0.41$, it is not as prominent as in the other networks. This hints that the network may be less *spatial* in nature, as is corroborated in our other experiments.

Link Prediction

We first evaluate our model by performing link prediction using 10-fold cross validation with a $90/10$ split for training and testing (i.e. 90% of the links are used for training the model and the remaining 10% are predicted) over each of the spatial networks. We compute the link predictions with our model in two different manners: (i) the predictive link probability and (ii) the maximum a-posterior (MAP) parameter configuration of the model. The predictive link probability, given in Eq. 6, is defined by integrating over the posterior probabilities of the model parameters to compute the probability of a link existing.

$$p(A_{ij}|D_{ij},k_i,k_j) = \int_{\alpha,\gamma,r_i,r_j} p(A_{ij},r_i,r_j,\alpha,\gamma|D_{ij},k_i,k_j)d\alpha d\gamma dr_i dr_j \quad (6)$$

Whereas using the MAP configuration simply requires plugging in the set of parameters that maximized the posterior probability. More formally, the MAP link prediction is given as follows:

$$p(A_{ij}|D_{ij},k_i,k_j) = p(A_{ij}|r_i^*,r_j^*,D_{ij},k_i,k_j,\alpha^*,\gamma^*) \quad (7)$$

$$\{r_i^*,r_j^*,\alpha^*,\gamma^*\} = argmax\ p(A_{ij},r_i,r_j,\alpha,\gamma|D_{ij},k_i,k_j)$$

where the node degrees, k_i and k_j, are computed from the observed network (i.e. held out links are not counted). Note that as we are predicting *new* links given an observed portion of the network, the actual degree values may change. Thus we do not constrain the model by enforcing that $k_i = \sum_{j \neq i} A_{ij}$.

Both of these methods consistently gave similar predictions, thus we only show results using the predictive link probability. To provide a baseline, we compare our model to (i) the configuration model $(k_i k_j / \sum_z k_z)$ (PA), (ii) exponential distance decay (ExpDist) [15,18], and (iii) empirical distance decay (EmpDist) [17]. To perform link prediction using these methods, we compute the expectation of an edge for each pair of nodes using the statistics collected from the training links. Because the normalizations used in each of these methods is based on the total number of links in the network, the expectation may result in values larger than 1. These values are thresholded and simply taken to be 1.

To evaluate the link prediction quality of the different methods, we employ area under the receiver operating characteristics (ROC) curve (see [60] for more details). Figure 7 shows the area under the ROC curve (AUC) aggregated over the 10-folds for each dataset. From these results, we notice several interesting trends. First, the configuration model (PA) (i.e. completely ignoring space) performs surprisingly well, with AUC values typically over 75%. Thus, while space certainly plays an important role in the formation of links in these datasets, node popularity is certainly an influential factor in determining network topology which must be taken into consideration. Second, *EmpDist* consistently outperforms both *PA* and *ExpDist*. Additionally, *ExpDist* performs only marginally better than *PA*, except for in the *C. elegans* network where it actually has worse performance. This is likely due to the fact that the true link distance distributions is not actually exponential, as we showed in our earlier analysis.

Lastly, *Radius* typically achieves better predictions than *EmpDist*, though with much higher variability (over the 10-folds). This is intuitive, since the radii provide more flexibility at the cost of additional model variables which need to be inferred. By accounting for additional community structure within the networks, *Radius+Comms*, provides a substantial improvement over *Radius* in all of the networks. In all of the networks except Internet,

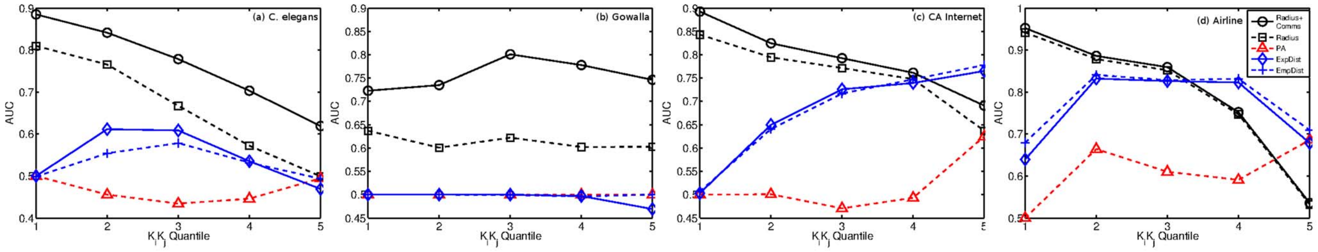

Figure 9. AUC measured over separate quantiles of the test data, split by the combined degrees of the nodes for which a link is being predicted, $k_i k_j$. The quantiles are shown on the x-axis, where 1 contains all node-pairs in which both nodes have low degree and 5 contains those in which both nodes have very high degrees.
doi:10.1371/journal.pone.0071293.g009

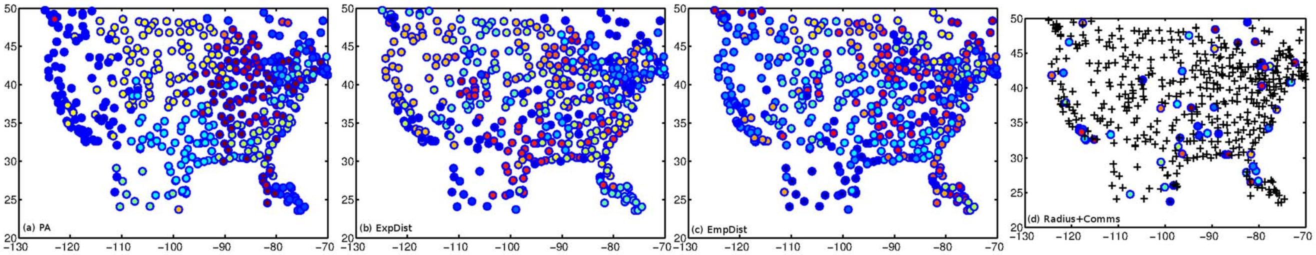

Figure 10. The communities detected by the different methods in the Airline network (best viewed in color). The communities identified by PA show a strong spatial structure, which is mostly maintained in *ExpDist* and *EmpDist* as well, although nodes on the fringe may switch to neighboring communities. In contrast, *Radius+Comms* identifies much fewer, though much more strongly integrated communities (nodes not belonging to any community are shown as black+'s) for which it is difficult to identify any real spatial structure.
doi:10.1371/journal.pone.0071293.g010

Table 3. Agreement between community detection methods.

C. elegans					
	Radius	**PA**	**ExpDist**	**EmpDist**	**Radius+Comm**
Radius		0.554	0.598	0.585	0.691
PA	0.554		0.629	0.699	0.538
ExpDist	0.598	0.629		0.693	0.533
EmpDist	0.585	0.699	0.693		0.525
Radius+Comm	0.691	0.538	0.533	0.525	
Gowalla					
	Radius	**PA**	**ExpDist**	**EmpDist**	**Radius+Comm**
Radius		0.961	0.737	0.972	0.339
PA	0.927		0.737	0.973	0.321
ExpDist	0.942	0.940		0.740	0.436
EmpDist	0.941	0.950	0.945		0.321
Radius+Comm	0.961	0.935	0.927	0.934	
CA Internet					
	Radius	**PA**	**ExpDist**	**EmpDist**	**Radius+Comm**
Radius		0.453	0.570	0.556	0.089
PA	0.791		0.444	0.482	0.099
ExpDist	0.856	0.786		0.531	0.092
EmpDist	0.846	0.803	0.884		0.088
Radius+Comm	0.881	0.802	0.870	0.880	
Airline					
	Radius	**PA**	**ExpDist**	**EmpDist**	**Radius+Comm**
Radius		0.542	0.609	0.646	0.140
PA	0.712		0.493	0.542	0.132
ExpDist	0.790	0.663		0.639	0.150
EmpDist	0.829	0.750	0.897		0.144
Radius+Comm	0.882	0.751	0.852	0.885	

The top triangular matrix contains normalized mutual information (NMI) scores comparing the resulting communities between the different methods. The bottom triangular matrix shows NMI over just the subset of nodes that *Radius+Comms* placed into a community. The number of nodes considered for each network were: (C. elegans) 277, (Gowalla) 134, (CA Internet) 28, (Airline)) 36. The first four rows (columns) are computed by using the referenced model as the null model and applying modularity optimization [40]. The last row (column), with the blue tinted background, is the result of our *Radius+Comms* model, in which the community structure is identified within the model itself.
doi:10.1371/journal.pone.0071293.t003

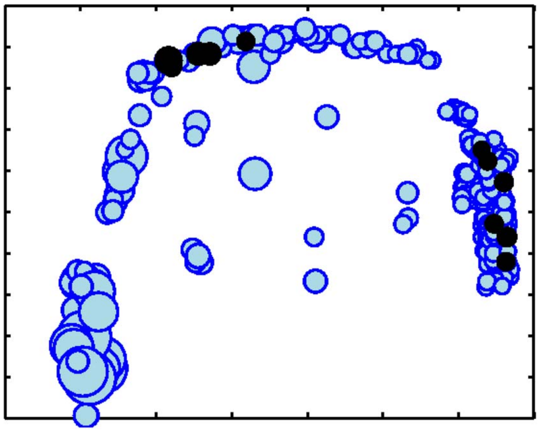

Figure 11. Nodes from the *C.elegans* neuronal network shown in their original positions. Sample communities identified by *Radius+Comms* are shown as black nodes.
doi:10.1371/journal.pone.0071293.g011

we also notice that *Radius+Comms* has much lower variance in its AUC (over the different folds) than *Radius*. This can be attributed to the fact that pairs of nodes between which a link was uncertain in the *Radius* model are likely to be fixed by adding these nodes to the same community, thus explaining part of the link structure more robustly. The high variance in the Internet network is the result of few communities being detected. We investigate the resulting communities in more depth in section.

Next, we break down the links according to distance and node degrees to further understand our model's performance. We split the test data into 5 quantiles based on pairwise node distance and degree, then compute the AUC over each quantile. The quantiles are computed such that there is an even split of links (i.e. true positives) in the testing data into each bin. Figures 8 and 9 show our results for splits based on pairwise distance and node degree respectively.

Comparing the methods by pairwise distance shows that the *Radius* and *Radius+Comms* models consistently provide higher AUC scores. The only surprise comes from the *C. elegans* and Internet networks at the largest distances, where *Radius* declines while *PA* and *EmpDist* both improve. Because PA improves in this quantile, it suggests that these links may be explained by the node popularity alone. Whereas the *Radius* model is putting too much weight on the distance between these nodes, the other models, with much weaker spatial components, capture these connections due to the popularity of the nodes. The shortcomings in the *Radius* model seem to be overcome in *Radius+Comms*, because the added community variables are able to help explain long distance connections.

Splitting the test data by combined node degrees shows an interesting trend in that all of the previous (global) models are universally bad at predicting edges between nodes with low degrees. This is because the primary source of information used for link prediction in these models is the node degree. Thus if a node is observed as having few connections, it is unlikely to have any more connections. In contrast, the *Radius* model encapsulates information about the network structure local to each node, which is critical to providing accurate predictions for these nodes. For example, if a node is observed to have only one connection but is in a region of low density (i.e. there are few nodes nearby), then any connection made with this node will be further away than the same node in a region of higher density. Whereas the other methods employ a global function of distance which would penalize this node for making such a connection, the radius in our model captures that this is normal given the node's surroundings.

The amount of improvement in link prediction quality our models achieve on low-degree nodes is especially promising. Due to the fact that many nodes are likely to have low degrees (since many networks follow the power-law degree distribution) and network structure alone provides very little information about these nodes, our modeling approach offers a substantial advantage over other techniques. Furthermore, these results emphasize the importance of accurately modeling the link-distance cost function.

Community Detection

In this section, we investigate the applicability of our models to the task of community detection in spatial networks. We compare the resulting communities identified by our *Radius+Comms* model with previous methods [15,17]. Additionally, we also use the *Radius* model as a the null comparison within modularity optimization [40]. Since no ground truth exists for the community structure in these networks, we provide a pairwise comparison of the different methods. We measure the consistency of the resulting communities across all of the different methods using normalized mutual information (NMI) [61]. By analyzing the similarity of the identified community structures, we show that our proposed model, *Radius+Comms*, captures only the very strongly connected groups of nodes. These are the communities which persist, despite the differences in the clustering objective functions (or the null models).

We observe that all of the spatial, modularity-based models tend to produce results more similar to each other than to the basic PA null model. This is intuitive, as each of these models is considering the same additional information about network structure, though they are incorporating this information differently. Additionally, the two baseline spatial null models, *ExpDist* and *EmpDist*, show similar levels of agreement amongst themselves indicating that even relatively small changes in the null model can force nodes on the fringe of a community to switch to another group. This is shown visually in figure 10.

In general, we see very little agreement between the communities discovered using the modularity-based approaches and *Radius+Comms*. This is due to two major differences in the objective function. First, modularity only optimizes within cluster edges and does not explicitly penalize strong connections between clusters. This is in contrast to our method which equally rewards within cluster links as well as penalizes between cluster links. Second, modularity forces all nodes to be placed into a cluster, whereas *Radius+Comms* contains a special *don't care* group for which nodes are unaffected by community structure. This provides additional modeling flexibility in that we can both find instances where community structure helps explain link structure as well as instances where nodes do not appear to be affected (i.e. link structure can be explained by distance and popularity alone).

However, examining the subset of nodes which are explicitly placed into communities in *Radius+Comms*, we find very strong agreement across all of the clustering methods (bottom half of tables ineach section in Table 3). The fact that much of the community structure found using our method persists even when the clustering objective function is modified, indicates that *Radius+Comms* is identifying only the most significant communities. In fact, the importance of the identified community structure is orated by our link prediction results as well. *Radius+Comms* offers substantial improvements over *Radius* in our ability to explain the network structure, and thus predict missing links across all of the data sets.

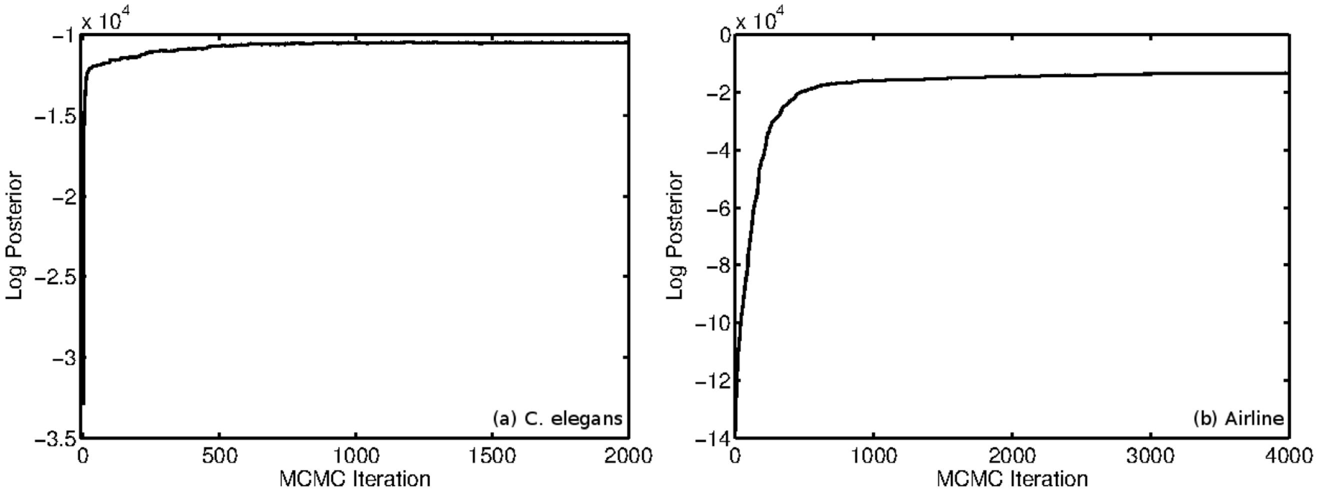

Figure 12. Log posterior trace plots from initialization run of the (a) *C. elegans* network and the (b) US Airline network. For *C. elegans*, we observe fast convergence of the log posterior in under 2,000 iterations, whereas for the Airline network, we observe the posterior is still rising, at a very slow rate, past 4,000 iterations.
doi:10.1371/journal.pone.0071293.g012

Upon further inspection, we see that the communities identified by *Radius+Comms* are in fact spatial anomalies. One such example of this is in the Airline network where we find that the Lake Charles Regional Airport in Lake Charles, Louisiana and the Chris Hadfield Airport in Sarnia, Ontario which are placed into the same community. These two airports are separated by more than 1,700 km, and the airports have a total of 2 and 1 recorded connections respectively. Given the size of these airports and the large distance separating them, such a connection is truly not expected.

Similarly, figure 11 shows an example communities identified in the *C. elegans* network. Despite being spatially diverse, the community is composed of functionally similar neurons. The community includes Ventral cord motor neurons and interneurons which play a role in locomotion. The functions of these neurons all surround the task of locomotion as well as collision detection [62,63]. These examples indicate that there is indeed a reasonable level of coherence within the communities.

Inference Analysis

Lastly, we discuss the convergence and mixing properties of our MCMC algorithm and provide a brief analysis of how the prior distributions influence the inference algorithm. To guarantee good mixing and quick convergence, we wish to provide a good initialization of the parameters. For each network, we run a short Markov chain and use the maximum a-posterior (MAP) configuration from that run to initialize the model parameters. While we find that we are able to converge quickly for most of the datasets, convergence on the airline network was particularly slow. We observe a large initial jump in the log posterior after the first few iterations when we move from the randomly initialized parameter values into a more coherent configuration.

However, unlike the other networks in which the log-posterior flattens out indicating that we have reached the mode of the distribution, the airline network slowly improves over several thousand iterations until it finally converges into a posterior mode. Such a slow convergence indicates that the posterior distribution may be rather diffuse for the given data and thus several parameter configurations may provide similarly adequate fits for the network. Figure 12 shows the log posterior from the *C. elegans* and US Airline networks. Despite the slow convergence on the Airline network, we still see consistent results across multiple runs.

Next, we investigate how the prior distributions on the model parameters influence the inference algorithm and model quality. We test this on synthetically generated networks to test how changing the prior parameters influences accuracy of the inferred

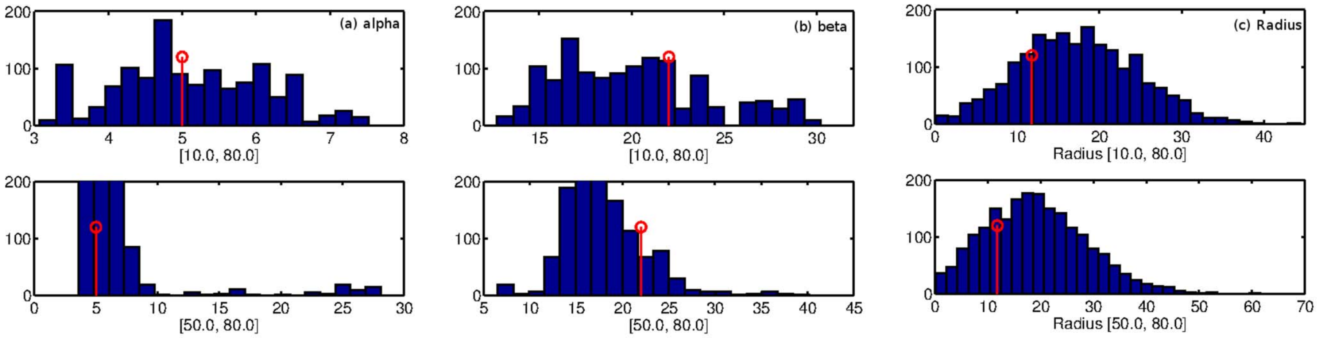

Figure 13. Comparison of posterior distributions under different settings of the prior parameters (run on synthetic data). The top row results from the prior $\mathcal{N}(10,80)$, and the bottom row uses $\mathcal{N}(50,80)$.
doi:10.1371/journal.pone.0071293.g013

Table 4. Generic approach for setting the variance of the prior distribution over model parameters.

Variable	Prior Mean	Prior Standard Deviation
r	$median(D(i,j)\|A_{ij}=1)$	$0.5std(D(i,j)\|A_{ij}=1)$
α	$(\mathcal{D}_{max}-\mu_r)/100$	$(\mathcal{D}_{max}-\mu_r)/100$
β	$(\mathcal{D}_{max}-\mu_r)/100$	$(\mathcal{D}_{max}-\mu_r)/100$
γ	1	100

doi:10.1371/journal.pone.0071293.t004

parameter values. We generated 10 synthetic networks using *Radius+Comms* model's generative process (after distributing nodes uniformly over a given region of space) so that we know the true parameter values. Then, we ran our inference algorithm on the observed networks using different settings for the prior distributions. Figure 13 shows the resulting posterior distributions, as well as the generating parameter values, for one synthetic network.

For all parameters, the top and bottom rows show the posterior distribution when the prior mean was set to 10 and 50 respectively. The prior variance was kept at 80 to capture our prior uncertainty in these parameters. For both settings of the prior, we see that all of the posteriors are centered around the the parameter value with which the observed networks were generated. We do notice a rather slight shift in the posterior when the prior mean was set to 50, though the posterior mode still converges to the correct area. In our experiments on the real networks, we have noticed that modifying the prior parameters has very little effect on the resulting posterior distributions, as corroborated in this simple analysis.

In our experiments on the real networks, table 4 describes a generic method used to set the prior mean and standard deviation for each variable based on the ROI and basic statistics of the networks. The prior for the radius is set to be a function of the median distance of linked nodes, so as to not over estimate the effect of the radius. The prior parameters for the global scaling variables, α and β are functions of the maximum distance between a pair of linked nodes, $\mathcal{D}_{max}=max(D(i,j)|A_{ij}=1)$. Notice that the prior standard deviations are set to be rather high to encode our level of uncertainty in the actual parameter values, allowing the data to have a strong influence over the infered values. Lastly, γ, is simply set to $\mathcal{N}(1,100)$ because the scale of this term is dependent on the pairwise distances.

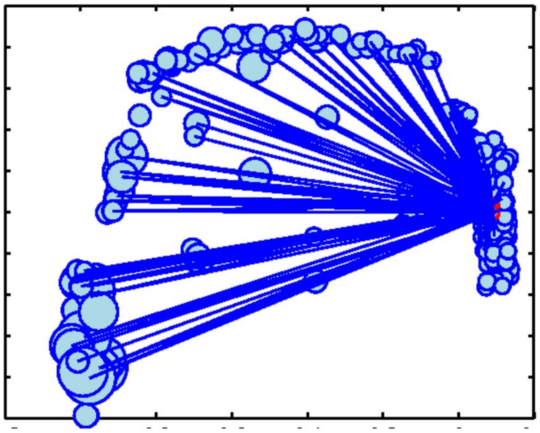

Figure 15. Connections formed by the AVA neurons (shown in red).
doi:10.1371/journal.pone.0071293.g015

We experimented with varying the prior parameters on the real and synthetic networks and found the results to remain consistent across trials, so all of our reported experiments set prior parameters according to table 4.

Discussion

In the previous section, we showed that our proposed models provide an accurate fit to several real world spatial networks. Next, we analyze the inferred parameter values for *Radius+Comms* on the *C. elegans* network. We focus on *C. elegans* because detailed information about the nodes (i.e. neurons) is available, thus we are best able to interpret and explain our findings [64].

The neurons with the largest radii are PVC[L/R] and DV[A/B]. The DVA neuron functions in mechanosensory integration, providing input to both the anterior and posterior touch circuits [64]. Neurons taking part in such sensory integration naturally need to interact with a wide variety of spatially disperse neurons in order to collect this information, thus explaining the need for a large spatial reach. The PVC[L/R] neurons are known to form

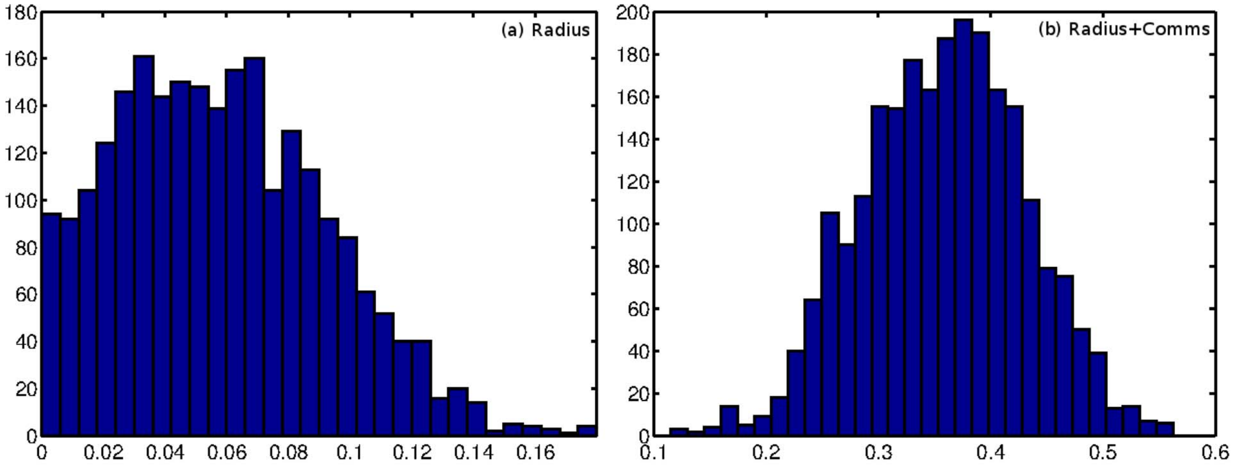

Figure 14. Posterior samples of the radius for the neuron PVCL, which has one of the largest (posterior average) radii in the network (in both models).
doi:10.1371/journal.pone.0071293.g014

synapses with the VB group of neurons (motor neurons) which are located in the head of the worm, as well as the DB neurons (dorsal motor neurons) which are located throughout the body of the worm. Given that the PVC[L/R] neurons are located in the tail, they must extend a long distance to form these links. We show histograms of the posterior distribution of the radius of PVCL for each of the models in figure 14.

The smallest radii belong to the AVE[L/R] and AVA[L/R] neurons, all of which are located in the head of the worm. Interestingly, it is known that the processes (axons and dendrites) of the AVE[L/R] neurons are restricted to the area above the vulva, which is typically found near the center of the worm body [62,64]. This limited spatial reach, combined with the fact that the neurons lie in the head of the worm, where neurons are most dense, explain this node's small radius. In contrast, the AVA[L/R] neurons are the pair with the largest degrees, with 76 and 74 connections respectively. Moreover, these neurons run the entire length of the ventral nerve cord as they function in forward and backward movement [62,64]. Given the wide reach of these neurons, it seems peculiar that they would not have larger radii. However, upon further inspection, we see that although they form many connections with neurons spread throughout the body of the worm, they also neglect to form connections with many neurons in the head (see figure 15). Because there is a high density of neurons in the head of the worm, if these neurons do not form connections with other neurons in this region, their radii will be penalized heavily. Thus, many neurons in this area have very small *spatial reach* and other nodes in less dense regions are forced to increase their *spatial reach* to pick up the slack.

Conclusions

We have introduced a novel, node-centric approach for modeling the link-distance cost function of a spatial network. To learn this function, we attach a latent radius parameter to each node which summarizes the local network structure. The radius parameter is influenced by the local density of surrounding, as well as the number of connections and their associated link distances. Additionally, we have provided a natural extension to this model which captures salient community structure, which cannot be explained due to spatial or node popularity effects.

We have shown experimentally that our models, *Radius* and *Radius+Comms*, result in higher quality link predictions across four different real-world spatial networks than competing techniques. Interestingly, the most substantial improvements came from predicting links between nodes with low observed degrees. That is, the nodes from which the network structure provides the least amount of information. Furthermore, we analyze the model parameters and offer interpretations of the inferred values on the test networks.

Studying the role of space in networks is critical to further our understanding of complex systems. In this work, we have introduced a model which offers the flexibility required to appropriately account for complicated link-distance cost functions as well as other connection properties. Our model provides a node-centric view of the unobserved link-distance cost function which influences the network structure. This approach offers greater modeling flexibility, and, as we have demonstrated, a more accurate representation of the data.

Acknowledgments

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notice herein.

Author Contributions

Conceived and designed the experiments: NL AS BR. Performed the experiments: NL. Analyzed the data: NL. Contributed reagents/materials/analysis tools: NL. Wrote the paper: NL BR AS.

References

1. Airoldi E, Goldenberg A, Zheng A, Fienberg S (2009) A survey of statistical network models. Machine Learning 2: 129–233.
2. Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: WWW.
3. Wong LH, Pattison P, Robins G (2006) A spatial model for social networks. Physica A: Statistical Mechanics and its Applications 360: 99–120.
4. Albert R, Barabási A (2000) Topology of evolving networks: local events and universality. Physical review letters 85: 5234–5237.
5. Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286: 509–512.
6. Fortuna M, Gómez-Rodríguez C, Bascompte J (2006) Spatial network structure and amphibian persistence in stochastic environments. Proceedings of the Royal Society B: Biological Sciences 273: 1429–1434.
7. Olesen J, Bascompte J, Dupont Y, Jordano P (2007) The modularity of pollination networks. Proceedings of the National Academy of Sciences 104: 19891.
8. Daraganova G, Pattison P, Koskinen J, Mitchell B, Bill A, et al. (2011) Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. Social Networks.
9. Ferretti L, Cortelezzi M (2011) Preferential attachment in growing spatial networks. Physical Review E 84: 016103.
10. Hayashi Y (2006) A review of recent studies of geographical scale-free networks. IPSJ Digital Courier 2: 155–164.
11. Gastner M, Newman M (2006) Optimal design of spatial distribution networks. Physical Review E 74: 016117.
12. Guimera R, Amaral L (2004) Modeling the world-wide airport network. The European Physical Journal B-Condensed Matter and Complex Systems 38: 381–385.
13. Metcalf S, Paich M (2005) Spatial Dynamics of Social Network Evolution. In: System Dynamics Society.
14. Barthélemy M (2011) Spatial networks. Physics Reports 499: 1–101.
15. Cerina F, De Leo V, Barthelemy M, Chessa A (2012) Spatial correlations in attribute communities. PLoS ONE 7: e37507.
16. Daraganova G, Pattison P, Koskinen J, Mitchell B, Bill A, et al. (2012) Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. Social Networks 34: 6–17.
17. Expert P, Evans T, Blondel V, Lambiotte R (2011) Uncovering space-independent communities in spatial networks. Proceedings of the National Academy of Sciences 108: 7663.
18. Yook S, Jeong H, Barabási A (2002) Modeling the internet's large-scale topology. Proceedings of the National Academy of Sciences 99: 13382–13386.
19. Hoff P, Raftery A, Handcock M (2002) Latent space approaches to social network analysis. Journal of the American Statistical Association 97: 1090–1098.
20. Hoff P (2009) Multiplicative latent factor models for description and prediction of social networks. Computational & Mathematical Organization Theory 15: 261–272.
21. Barabási A (2012) Network science: Luck or reason. Nature 489: 507–508.
22. Erdős P, Rényi A (1960) On the evolution of random graphs. Magyar Tud Akad Mat Kutató Int Közl 5: 17–61.
23. Newman M (2003) The structure and function of complex networks. SIAM review: 167–256.
24. Watts D, Strogatz S (1998) Collective dynamics of small-world networks. Nature 393: 440–442.
25. White D, Kejžar N, Tsallis C, Farmer D, White S (2006) Generative model for feedback networks. Physical Review E 73: 016119.
26. Waxman B (1988) Routing of multipoint connections. IEEE Journal on Selected Areas in Communications 6: 1617–1622.
27. Kaiser M (2004) Spatial growth of real-world networks. Phys Rev E 69: 036103.
28. Dall J, Christensen M (2002) Random geometric graphs. Physical Review E 66: 016121.
29. Penrose M (2003) Random geometric graphs, volume 5. Oxford, UK: Oxford University Press.

30. Barthélemy M (2003) Crossover from scale-free to spatial networks. EPL (Europhysics Letters) 63: 915.
31. Kosmidis K, Havlin S, Bunde A (2008) Structural properties of spatially embedded networks. EPL (Europhysics Letters) 82: 48005.
32. Barrat A, Barthélemy M, Vespignani A (2005) The effects of spatial constraints on the evolution of weighted complex networks. Journal of Statistical Mechanics: Theory and Experiment 2005: P05003.
33. Xulvi-Brunet R, Sokolov I (2002) Evolving networks with disadvantaged long-range connections. Physical Review E 66: 026118.
34. Barnett L, Di Paolo E, Bullock S (2007) Spatially embedded random networks. Physical Review E 76: 056115.
35. Turgut D, Atilgan A, Atilgan C (2010) Assortative mixing in close-packed spatial networks. PLoS One 5: e15551.
36. Kaiser M, Hilgetag C (2004) Modeling the development of cortical systems networks. Neurocomputing 58: 297–302.
37. Bullock S, Barnett L, Di Paolo E (2010) Spatial embedding and the structure of complex networks. Complexity 16: 20–28.
38. Voges N, Aertsen A, Rotter S (2007) Statistical analysis of spatially embedded networks: From grid to random node positions. Neurocomputing 70: 1833–1837.
39. Wilson A (1967) A statistical theory of spatial distribution models. Transp Res: 253–269.
40. Newman M (2006) Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103: 8577–8582.
41. Lennartsson J, Håkansson N, Wennergren U, Jonsson A (2012) Specnet: A spatial network algorithm that generates a wide range of specific structures. PLoS ONE 7: e42679.
42. Newman M (2002) Assortative mixing in networks. Physical Review Letters 89: 208701.
43. Handcock M, Raftery A, Tantrum J (2007) Model-based clustering for social networks. Journal of the Royal Statistical Society: Series A (Statistics in Society) 170: 301–354.
44. Krivitsky P, Handcock M, Raftery A, Hoff P (2009) Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. Social networks 31: 204–213.
45. Sarkar P, Chakrabarti D, Moore A (2010) Theoretical justification of popular link prediction heuristics. In: International Conference on Learning Theory (COLT). pp. 295–307.
46. Hoff P (2005) Bilinear mixed-effects models for dyadic data. Journal of the American Statistical Association 100: 286–295.
47. Hoff P (2007) Modeling homophily and stochastic equivalence in symmetric relational data. In: Neural Information Processing Systems.
48. Snijders T (2002) Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure 3: 1–40.
49. Wasserman S, Pattison P (1996) Logit models and logistic regressions for social networks: An introduction to markov graphs and p. Psychometrika 61: 401–425.
50. Airoldi E, Blei D, Fienberg S, Xing E (2008) Mixed membership stochastic blockmodels. The Journal of Machine Learning Research 9: 1981–2014.
51. Karrer B, Newman M (2011) Stochastic blockmodels and community structure in networks. Physical Review E 83: 016107.
52. Nowicki K, Snijders T (2001) Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association 96: 1077–1087.
53. Boguná M, Krioukov D, Claffy K (2008) Navigability of complex networks. Nature Physics 5: 74–80.
54. Boguná M, Papadopoulos F, Krioukov D (2010) Sustaining the internet with hyperbolic mapping. Nature Communications 1: 62.
55. Krioukov D, Papadopoulos F, Vahdat A, Boguná M (2009) Curvature and temperature of complex networks. Physical Review E 80: 035101.
56. Krioukov D, Papadopoulos F, Kitsak M, Serrano M, Boguná M (2012) Popularity versus similarity in growing networks. Bulletin of the American Physical Society 57.
57. Serrano M, Krioukov D, Boguná M (2008) Self-similarity of complex networks and hidden metric spaces. Physical review letters 100: 78701.
58. Bonahon F (2009) Low-dimensional geometry: From Euclidean surfaces to hyperbolic knots, volume 49. Amer Mathematical Society.
59. Datasets. Available: http://dx.doi.org/10.6084/m9.figshare.153828.
60. Fawcett T (2006) An introduction to roc analysis. Pattern recognition letters 27: 861–874.
61. Estévez P, Tesmer M, Perez C, Zurada J (2009) Normalized mutual information feature selection. Neural Networks, IEEE Transactions on 20: 189–201.
62. Riddle D, Blumenthal T, Meyer B, Priess J (1997) C. elegans II, volume 33. CSHL press.
63. Varshney L, Chen B, Paniagua E, Hall D, Chklovskii D (2011) Structural properties of the *Caenorhabditis elegans* neuronal network. PLoS Computational Biology 7: e1001066.
64. Worm atlas. Available: http://www.wormatlas.org/. Accessed: 2012 Oct 9.
65. Diggle P (2003) Statistical Analysis of Spatial Point Patterns. Mathematics in biology. Hodder Arnold. Available: ttp://books.google.com/books?id = fnFhQgAACAAJ.