bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Prioritizing network communities

Marinka Zitnik,[1] Rok Sosič,[1] and Jure Leskovec[1,2,*]

[1] Computer Science Department, Stanford University, Stanford, CA, USA
[2] Chan Zuckerberg Biohub, San Francisco, CA, USA

[*] To whom correspondence should be addressed; E-mail: jure@cs.stanford.edu

Uncovering modular structure in networks is fundamental for advancing the understanding of complex systems in biology, physics, engineering, and technology. Community detection provides a way to computationally identify candidate modules as hypotheses, which then need to be experimentally validated. However, validation of detected communities requires expensive and time consuming experimental methods, such as mutagenesis in a wet biological laboratory. As a consequence only a limited number of communities can be experimentally validated, and it is thus important to determine which communities to select for downstream validation and experimentation. Here we develop CRANK, an automatic method for prioritizing network communities and identifying the most promising ones for further experimentation. CRANK efficiently evaluates robustness and magnitude of structural features of each community and then combines these features to obtain the community prioritization. CRANK can be used with any community detection method and scales to large networks. It needs only information provided by the network structure and does not require any additional metadata or labels. However, when available, CRANK can incorporate domain-specific information to further boost performance. Experiments on many diverse and important biological networks demonstrate that the proposed approach effectively prioritizes communities, yielding a near 50-fold improvement in community prioritization over a baseline ordering of detected communities. Taken together, CRANK represents a network-based approach to identify high-quality communities, even for domains at the frontier of science where supervised meta information is not available.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Networks exhibit modular structure[1] and uncovering it is fundamental for advancing the understanding of complex systems across sciences[2,3]. Methods for community detection[4], also called node clustering or graph partitioning, allow for computational detection of modular structure by identifying a division of network's nodes into groups, also called communities[5–10]. Such communities provide predictions/hypotheses about potential modules of the network, which then need to be experimentally validated and confirmed. However, in large networks, community detection methods typically identify many thousands of communities[6,7] and only a small fraction can be rigorously tested and validated by follow-up experiments. For example, gene communities detected in a gene interaction network[11] provide predictions/hypotheses about disease pathways[2,3], but to confirm these predictions scientists have to test every detected community by performing experiments in a wet laboratory[3,8]. Because experimental validation of detected communities is resource-intensive and generally only a small number of communities can be investigated, one must prioritize the communities in order to choose which ones to investigate experimentally.

In the context of biological networks, several methods for community or cluster analysis have been developed[2,3,12–15]. However, these methods crucially rely and depend on knowledge in external databases, such as Gene Ontology (GO) annotations[16], protein domain databases, gene expression data, patient clinical profiles, and sequence information, in order to calculate the quality of communities derived from networks. Furthermore, they require this information to be available for all communities. This means that if genes in a given community are not present in a gene knowledge database then it is not possible for existing methods to even consider that community. This issue is exacerbated because knowledge databases are incomplete and biased toward better-studied genes[11]. Furthermore, these methods do not apply in domains at the frontier of science where domain-specific knowledge is scarce or non-existent, such as in the case of cell-cell similar-

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ity networks[17], microbiome networks[18,19], and chemical interaction networks[20]. Thus, there is a need for a general solution to prioritize communities based on network information only.

Here, we present CRANK, a general approach that takes a network and detected communities as its input and produces a ranked list of communities, where high-ranking communities represent promising candidates for downstream experiments. CRANK can be applied in conjunction with any community detection method (**SI Text** Sections S2 and S5) and needs only the network structure, requiring no domain-specific meta or label information about the network. However, when domain-specific supervised information is available, CRANK can integrate this extra information to boost performance (**SI Text** Sections S9 and S10). CRANK can thus prioritize communities that are well characterized in knowledge bases, such as GO annotations, as well as poorly characterized communities with limited or no annotations. Furthermore, CRANK is based on rigorous statistical methods to provide an overall rank for each detected community.

CRANK community prioritization methodology consists of the following steps (Figure 1): (a) find communities using an existing, preferred community detection method (Figure 1a); (b) compute for each community four CRANK defined community prioritization metrics, which capture key structural features of the community (Figure 1b); (c) combine the community metrics via a novel CRANK aggregation method into a single overall score for each community (Figure 1c); (d) prioritize communities by ranking them by their decreasing overall score (Figure 1d).

CRANK uses four different metrics to characterize network connectivity features for each detected community (see Methods). These metrics evaluate the magnitude of structural features as well as their robustness against noise in the network structure. The rationale here is that high priority communities have high values of metrics and are also stable with respect to network perturbations. If a small change in the network structure—an edge added here, another deleted there—

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

significantly changes the value of a prioritization metric then the community will not be considered high priority. We derive analytical expressions for calculating these metrics, which make CRANK computationally efficient and applicable to large networks (**SI Text** Section S2). Because individual metrics may have different importance in different networks, a key element of CRANK is a novel rank aggregation method. This method combines the values of the four metrics into a single score for each community, which then determines the community's rank (see Methods and **SI Text** Section S4). CRANK's aggregation method adjusts the impact of each metric on the ranking in a principled manner across different networks and also across different communities within a network, leading to robust rankings and a high-quality prioritization of communities (**SI Text** Section S5).

Results

Synthetic networks

We first demonstrate CRANK by applying it to synthetic networks with planted community structure (Figure 2a). The goal of community prioritization is to identify communities that are most promising candidates for follow-up investigations. Since communities provide predictions about the modular structure of the network, promising candidates are communities that best correspond to the underlying modules. Thus, in this synthetic example, the aim of community prioritization can be seen as to rank communities based on how well they represent the underlying planted communities, while only utilizing information about network structure and without any additional information about the planted communities. We quantify prioritization quality by measuring the agreement between a ranked list of communities produced by CRANK and the gold standard ranking. In the gold standard ranking, communities are ordered in the decreasing order of how accurately each community reconstructs its corresponding planted community.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We experiment with random synthetic networks with planted community structure (Figure 2a), where we use a generic community detection method[7] to identify communities and then prioritize them using CRANK. We observe that CRANK produces correct prioritization—using only the unlabeled network structure, CRANK places communities that better correspond to planted communities towards the top of the ranking (Figure 2b), which indicates that CRANK can identify accurately detected communities by using the network structure alone and having no other data about planted community structure. Comparing the performance of CRANK to alternative ranking techniques, such as modularity[5] and conductance[21], we observe that CRANK performs 149% and 37% better than modularity and conductance, respectively, in terms of the Spearman's rank correlation between the generated ranking and the gold standard community ranking (Figure 2c). Moreover, we observed no correlation with the gold standard ranking when randomly ordering the detected communities. Although zero correlation is expected, poor performance of random ordering is especially illuminating because prioritization of communities is typically ignored in current network community studies.

Networks of medical drugs with shared target proteins

Community rankings obtained by CRANK provide a rich source of testable hypotheses. For example, we consider a network of medical drugs where two drugs are connected if they share at least one target protein (Figure 3a). Because drugs that are used to treat closely related diseases tend to share target proteins[22], we expect that drugs belonging to the same community in the network will be rich in chemicals with similar therapeutic effects. Identification of these drug communities hence provides an attractive opportunity for finding new uses of drugs as well as for studying drugs' adverse effects[22].

After detecting drug communities using a standard community detection method[7], CRANK

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

relies only on the network structure to prioritize the communities. We evaluate ranking performance by comparing it to metadata captured in external chemical databases and not used by the ranking method. We find that CRANK assigns higher priority to communities whose drugs are pharmacogenomically more similar (Figure 3b), indicating that higher-ranked communities contain drugs with more abundant drug-drug interactions, more similar chemical structure, and stronger textual associations. In contrast, ranking communities by modularity score gives a poor correspondence with information in the external chemical databases (Figure 3c).

We observe that the top ranked communities are composed from an unusual set of drugs (Figure 3a and **SI Table**), yet drugs with unforeseen community assignment may represent novel candidates for drug repurposing[22]. Examining the highest ranked communities, we do not expect mifepristone, an abortifacient used in the first months of pregnancy, to appear together with a group of drugs used to treat inflammatory diseases. Another drug with unanticipated community assignment is minaprine, a psychotropic drug that is effective in the treatment of various depressive states[23]. Minaprine is an antidepressant that antagonizes behavioral despair; however, it shares target proteins with several cholinesterase inhibitors. Two examples of such inhibitors are physostigmine, used to treat glaucoma, and galantamine, a drug investigated for the treatment of moderate Alzheimer's disease[24]. In the case of minaprine, an antidepressant, it was just recently shown that this drug is also a cognitive enhancer that may halt the progression of Alzheimer's disease[25]. It is thus attractive that CRANK identified minaprine as a member of a community of primarily cholinesterase inhibitors, which suggests minaprine's potential for drug repurposing for Alzheimer's disease.

The analysis here was restricted to drugs approved for medical use by the U.S. Food and Drug Administration, because these drugs are accompanied by rich metadata that was used for

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

evaluating community prioritization. We find that when CRANK integrates drug metadata into its prioritization model, CRANK can generate up to 55% better community rankings, even when the amount of additional information about drugs is small (**SI Text** Section S10). However, approved medical drugs represent less than one percent of all small molecules with recorded interactions. Many of the remaining 99% of these molecules might be candidates for medical usage or drug repurposing but currently have little or no metadata in the chemical databases. This fact further emphasizes the need for methods such as CRANK that can prioritize communities based on network structure alone while not relying on any metadata in external chemical databases.

Gene and protein interaction networks

CRANK can also prioritize communities in molecular biology networks, covering a spectrum of physical, genetic, and regulatory gene interactions[11]. In such networks, community detection is widely used because gene communities tend to correlate with cellular functions and thus provide hypotheses about biological pathways and protein complexes[2,3].

CRANK takes a network and communities detected in that network, and produces a rank-ordered list of communities. As before, while CRANK ranks the communities purely based on network structure, the external metadata about molecular functions, cellular components, and biological processes is used to assess the quality of the community ranking.

Considering highest ranked gene communities, CRANK's ranking contains on average 5 times more communities whose genes are significantly enriched for cellular functions, components, and processes[16] than random prioritization, and 13% more significantly enriched communities than modularity- or conductance-based ranking (**SI Text**, Section S11.4). For example, in the human protein-protein interaction network, the highest ranked community by CRANK is composed of 20 genes, including *PORCN*, *AQP5*, *FZD6*, *WNT1*, *WNT2*, *WNT3*, and other members of

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the Wnt signaling protein family[26] (**SI Text**, Section S11.4). Genes in that community form a biologically meaningful group that is functionally enriched in the Wnt signaling pathway processes (p-value = 6.4×10^{-23}), neuron differentiation (p-value = 1.6×10^{-15}), cellular response to retinoic acid (p-value = 2.9×10^{-14}), and in developmental processes (p-value = 9.2×10^{-10}).

Functional annotation of molecular networks is largely unavailable and incomplete, especially when studied objects are not genes but rather other entities, *e.g.*, miRNAs, mutations, single nucleotide variants, or genomic regions outside protein-coding loci[27]. Thus it is often not possible to simply rank the communities by their functional enrichment scores. In such scenarios, CRANK can prioritize communities reliably and accurately using only network structure without necessitating any external databases. Gene communities that rank at the top according to CRANK represent predictions that could guide scientists to prioritize resource-intensive laboratory experiments.

Megascale cell-cell similarity networks

Single-cell RNA sequencing has transformed our understanding of complex cell populations[28]. While many types of questions can be answered using single-cell RNA-sequencing, a central focus is the ability to survey the diversity of cell types and composition of tissues within a sample of cells.

To demonstrate that CRANK scales to large networks, we used the single-cell RNA-seq dataset containing 1,306,127 embryonic mouse brain cells[29] for which no cell types are known. The dataset was preprocessed using standard procedures to select and filter the cells based on quality-control metrics, normalize and scale the data, detect highly variable genes, and remove unwanted sources of variation[9]. The dataset was represented as a weighted graph of nearest neighbor relations (edges) among cells (nodes), where relations indicated cells with similar gene expression patterns calculated using diffusion pseudotime analysis[30]. To partition this graph into highly in-

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

terconnected communities we apply a community detection method proposed for single-cell data[8]. The method separates the cells into 141 fine-grained communities, the largest containing 18,788 (1.8% of) and the smallest only 203 (0.02% of) cells. After detecting the communities, CRANK takes the cell-cell similarity network and the detected communities, and generates a rank-ordered list of communities, assigning a priority to each community. CRANK's prioritization of communities derived from the cell-cell similarity network takes less than 2 minutes on a personal computer.

In the cell-cell similarity network, one could assume that top-ranked communities contain highly distinct marker genes[31], while low-ranked communities contain marker genes whose expression levels are spread out beyond cells in the community. To test this hypothesis, we identify marker genes for each detected community. In particular, for each community we find genes that are differentially expressed in the cells within the community[9] relative to all cells that are not in the community.

We find that high-ranked communities in CRANK contain cells with distinct marker genes, confirming the above hypothesis (average z-score of marker genes with respect to the bulk mean gene expression was above 200 and never smaller than 150) (Figure 4a-b). In contrast, cells in low-ranked communities show a weak expression activity diffused across the entire network and no community-specific expression activity (Figure 4c-d). Examining cells assigned to the highest-ranked community (rank 1 community) in CRANK, we find that most differentially expressed genes are *TYROBP*, *C1QB*, *C1QC*, *FCER1G*, and *C1QA* (at least a 200-fold difference in normalized expression with respect to the bulk mean expression[9]). It is known that these are immunoregulatory genes and that they play important roles in signal transduction in dendritic cells, osteoclasts, macrophages, and microglia[32]. In contrast, low-ranked communities (Figure 4 visualizes rank 139, rank 140, and rank 141 communities) contain predominantly cells in which genes show

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

no community-specific expression. Genes in communities ranked lower by CRANK hence do not have localized mRNA expression levels, suggesting there are no good marker genes that define those communities[28]. Since the expression levels of mRNA are linked to cellular function and can be used to define cell types[28], the analysis here points to the potential of using highest-ranked communities in CRANK as candidates to characterize cells at the molecular level, even in datasets where no cells are yet classified into cell types.

Analysis of CRANK prioritization approach

The CRANK approach can be applied with any community detection method and can operate on directed, undirected, and weighted networks. Furthermore, CRANK can also use external domain-specific information to further boost prioritization performance (**SI Text** Section S10). Results on diverse biological, information, and technological networks and on different community detection methods show that the second best performing approach changes considerably across networks, while CRANK always produces the best result, suggesting that it can effectively harness the network structure for community prioritization (**SI Text** Section S8.1). CRANK automatically adjusts weights of the community metrics in the prioritization, resulting in each metric participating with different intensity across different networks (**SI Text** Figure S6). This is in sharp contrast with deterministic approaches, which are negatively impacted by heterogeneity of network structures and network community models employed by different community detection methods. The four CRANK community prioritization metrics are essential and complementary. CRANK metrics considered together perform on average 45% better than the best single CRANK metric, and 26% better than any subset of three CRANK metrics (**SI Text** Section S8.3). CRANK performs on average 38% better than approaches that combine alternative community metrics (**SI Text** Section S8.4). Furthermore, CRANK can easily integrate any number of additional and domain-specific

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

community metrics[2,12–15], and performs well in the presence of low-signal and noisy metrics (**SI Text** Section S9). Furthermore, CRANK outperforms alternative approaches that combine the metrics by approximating NP-hard rank aggregation objectives (**SI Text** section S8.5).

Discussion

The task of community prioritization is to rank-order communities detected by a community detection method such that communities with best prospects in downstream analysis are ranked towards the top. We demonstrated that prioritizing communities in biological, information, and technological networks is important for maximizing the yield of downstream analyses and experiments. Prior efforts crucially depend on external meta information to calculate the quality of communities with an additional constraint that this information has to be available for all communities. We devised a principled approach for the task of community prioritization. Although the approach does not need any meta information, it can utilize such information if it is available. Furthermore, CRANK is applicable even when the meta information is noisy, incomplete, or available only for a subset of communities.

The CRANK community ranking is based on the premise that high priority communities produce high values of community prioritization metrics and that these metrics are stable with respect to small perturbations of the network structure. Our findings support this premise and suggest that both the magnitude of the metrics and the robustness of underlying structural features have an important role in the performance of CRANK across a wide range of networks (**SI Text** Sections S8.2 and S8.4). CRANK can easily be extended using existing network metrics and can also consider new domain-specific scoring metrics (**SI Text** Sections S9 and S10). Thus, it would be especially interesting to apply it to networks, where rich meta information exists and interesting domain-specific scoring metrics can be developed, such as protein interaction networks with dis-

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ease pathway meta information[33], and molecular networks with genome-wide associations[34]. We believe that the CRANK approach opens the door to principled methods for prioritizing communities in large networks and, when coupled with experimental validation, can help us to speed-up scientific discovery process.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Community prioritization model

CRANK prioritizes communities based on the robustness and magnitude of multiple structural features of each community. For each feature f, we specify a corresponding prioritization metric r_f, which captures the magnitude and the robustness of f. Robustness of f is defined as the change in the value of f between the original network and its randomly perturbed version. The intent here is that high quality communities will have high values of f and will also be robust to perturbations of the network structure. We define and discuss specific prioritization metrics later. Here, we first present the overall prioritization model.

Random perturbations of the network are based on rewiring of α fraction of the edges in a degree preserving manner[35] (**SI Text** Section S2). Parameter α measures perturbation intensity; a value close to zero indicates that the network has only a few edges rewired whereas a value close to one corresponds to a maximally perturbed network, which is a random graph with the same degree distribution as the original network.

Even though the prioritization model is framed conceptually in terms of perturbing the network by rewiring its edges, CRANK never actually rewires the network when calculating the prioritization metrics. Network rewiring is a computationally expensive operation. Instead, we derive analytical expressions that evaluate the metrics in a closed form without physically perturbing the network (**SI Text** Section S2), which leads to a substantial increase in scalability of CRANK.

Given structural feature f, we define prioritization metric r_f to quantify the change in the value of f between the original and the perturbed network. We want r_f to capture the magnitude of feature f in the original network as well as the change in the value of f between the network and its perturbed version.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We define prioritization metric r_f for community C as:

$$r_f(C;\alpha) = \frac{f(C)}{1 + d_f(C,\alpha)}, \tag{1}$$

where $f(C)$ is the feature value of community C in the original network, α measures perturbation intensity, $d_f(C,\alpha) = |f(C) - f(C|\alpha)|$ is the change of the feature value for community C between the network and its α-perturbed version, and $f(C|\alpha)$ is the value of feature f in the α-perturbed version of the network.

Generally, higher priority communities will have higher values of r_f. In particular, as f can take values between zero and one, then r_f also takes values between zero and one. r_f attains value of zero for community C whose value of $f(C)$ is zero. When $f(C)$ is nonzero, then $r_f(C;\alpha)$ down-weights it according to the sensitivity of community C to network rewiring. $f(C)$ is down-weighted by the largest amount when it changes as much as possible under the network perturbation (*i.e.*, $d_f(C,\alpha) = 1$). And, $f(C)$ remains unchanged when community C is maximally robust to network perturbation (*i.e.*, $d_f(C,\alpha) = 0$).

Community prioritization metrics

Prioritization metric $r_f(C)$ captures the magnitude as well as the robustness of structural feature f of community C. We define four different community prioritization metrics r_f. Through empirical analysis we show that these metrics holistically and non-redundantly quantify different features of network community structure (**SI Text** Section S8). Each metric is necessary and contributes positively to the performance of CRANK. We combine these metrics into a global ranking of communities using a rank aggregation method that we describe later.

Given a network $G(\mathcal{V},\mathcal{E},\mathcal{C})$ with nodes $\mathcal{V}$, edges $\mathcal{E}$, and detected communities $\mathcal{C}$, CRANK can be applied in conjunction with any statistical community detection method that allows for com-

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

puting the following three quantities: (1) the probability of node u belonging to a given community C, $p_C(u) = p(u \in C)$, (2) the probability of an edge $p(u, v) = p((u, v) \in \mathcal{E})$, and (3) a contribution of community C towards the existence of an edge (u, v), $p_C(u, v) = p((u, v) \in \mathcal{E}|u, v \in C)$. Many commonly used community detection methods allow for computing the above three quantities (**SI Text** Section S5).

Our rationale in defining the prioritization metrics is to measure properties that determine a high quality community, which is also robust and stable with respect to small perturbations of the network. For example, a genuine high quality community should provide good support for the existence of edges between its members in the original network as well as in the perturbed version. If a small change in the network structure—an edge added here, another deleted there—can completely change the value of the prioritization metric then the community should not be considered high quality. Analogously, a high quality community should have low confidence for edges pointing outside of the community both in the original as well as in the perturbed network.

Community likelihood

The community likelihood metric quantifies the overall connectivity of a given community. It measures the likelihood of the network structure induced by the nodes in the community. Note that the metric does not simply count the edges but considers them in a probabilistic way. As such it quantifies how well the observed edges can be explained by the community C. The intuition is that high quality community will contribute a large amount of likelihood to explain the observed edges.

We formalize the community likelihood for a given community C as follows:

$$f_l(C|\alpha) = \prod_{u \in C} p_C(u) \prod_{v \in C} s_C(u, v|\alpha), \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where $s_C(u, v|\alpha)$ is defined as follows:

$$s_C(u, v|\alpha) = \begin{cases} p_C(v)p_C(u, v|\alpha) \text{ if } (u, v) \in \mathcal{E} \\ p_C(v)(1 - p_C(u, v|\alpha)) \text{ if } (u, v) \notin \mathcal{E}. \end{cases}$$

Here, $p_C(u, v|\alpha)$ is a contribution of community C towards the creation of edge (u, v) under network perturbation intensity α. We derive analytical expressions for $p_C(u, v|\alpha)$ which allows us to compute their values without ever actually perturbing the network (**SI Text** Section S2).

Here (and for the other three prioritization metrics) we evaluate the feature in the original network ($f_l(C) = f_l(C|\alpha = 0)$) as well as in the slightly perturbed version of the network ($f_l(C|\alpha = 0.15)$). We then combine the two scores using the prioritization metric formula in Eq. (1).

Community density

In contrast to community likelihood, which quantifies the contribution of a community to the overall edge likelihood, community density simply measures the overall strength of connections within the community. By considering edge probabilities that are not conditioned on the community C, density implicitly takes into consideration potentially hierarchical and overlapping community structures. When a community is nested inside other communities, these enclosing communities contribute to the increased density of community's internal edges.

Formally, we define the density of a community as the joint probability of the edges between community members. Assuming network perturbation intensity α, density of community C is defined as:

$$f_d(C|\alpha) = \prod_{\substack{(u,v)\in\mathcal{E} \\ u\in C, v\in C}} p(u, v|\alpha), \tag{3}$$

where $p(u, v|\alpha)$ is the probability of edge (u, v) under network perturbation intensity α. We derive

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

analytical expression for $p(u, v|\alpha)$ which allows us to compute their values without ever actually perturbing the network (**SI Text** Section S2).

Community boundary

To complement the internal connectivity measured by community density, community boundary considers the strength of edges leaving the community. A structural feature of a high quality community is its good separation from the surrounding parts of the network. In other words, a high quality community should have sharp edge boundary, *i.e.* $B_C = \{(u, v) \in \mathcal{E}; u \in C, v \notin C\}$[4]. This intuition is captured by accumulating the likelihood against edges connecting the community with the rest of the network:

$$f_b(C|\alpha) = \prod_{\substack{u \in C \\ v \in \mathcal{V} \setminus C}} (1 - p(u, v|\alpha)). \tag{4}$$

The evaluation of Eq. (4) takes computational time linear in the size of the network, which is impractical for large networks with many detected communities. To speed up the calculations, we use negative sampling (**SI Text** Section S2) to calculate the value of Eq. (4), and thereby reduce the computational complexity of the boundary metric to time that depends linearly on the number of edges leaving the community.

Community allegiance

Last we introduce community allegiance. We define community allegiance as the preference for nodes to attach to other nodes that belong to the same community. Allegiance measures the fraction of nodes in a community for which the total probability of edges pointing inside the community is larger than probability of edges that point to the outside of the community. For a given community

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

C and network perturbation intensity α, community allegiance is defined as:

$$f_a(C|\alpha) = \frac{1}{|C|} \sum_{u \in C} \delta(\sum_{v \in N_u \cap C} p(u,v|\alpha) \geq \sum_{v \in N_u \setminus C} p(u,v|\alpha)), \tag{5}$$

where N_u is a set of network neighbors of u and δ is the indicator function, $\delta(x) = 1$ if x is true, and $\delta(x) = 0$, otherwise.

Community has high allegiance if nodes in the community tend to be more strongly connected to other members of the community than to the rest of the network. In a community with no significant allegiance this metric takes a value that is close to zero or changes substantially when the network is only slightly perturbed. However, in the presence of substantial community allegiance, the metric takes large values and is not sensitive to edge perturbation.

Combining community prioritization metrics

We just defined four community prioritization metrics: likelihood, density, boundary, and allegiance. Each metric on its own provides a useful signal for prioritizing communities (**SI Text** Section S8). However, scores of each metric might be biased, have high variance, and behave differently across different networks (Figure S6). It is thus essential to combine the values of individual metrics into a single aggregated score.

We develop an iterative unsupervised rank aggregation method that, without requiring an external gold standard, combines the prioritization metrics into a single aggregated prioritization of communities. The method is outlined in Figure . It naturally takes into consideration the fact that importance of individual prioritization metrics varies across networks and across community detection methods. The aggregation method starts by representing the values of each prioritization metric with a ranked list. In each ranked list, communities are ordered by the decreasing value of the metric. The method then determines the contribution of each ranked list to the aggregate

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

prioritization by calculating importance weights. The calculation is based on Bayes factors[36–38], an established tool in statistics. Each ranked list has associated a set of importance weights. Importance weights can vary with rank in the list. The method then calculates the aggregated prioritization of communities in an iterative manner by taking into account uncertainty that is present across different ranked lists and within each ranked list.

To calculate the weights without requiring gold standard, the method uses a two-stage iterative procedure. After initializing the aggregated prioritization, the method alternates between the following two stages until no changes in the aggregated prioritization are observed: (1) use the aggregated prioritization to calculate the importance weights for each ranked list, and (2) re-aggregate the ranked lists based on the importance weights calculated in the previous stage.

The model for aggregating community prioritization metrics, the algorithm, and the analysis of its computational time complexity are detailed in **SI Text** Section S4 and Section S5. The complete algorithm of CRANK approach is provided in **SI Text** Section S5.

Code and data availability. All relevant data are public and available from the authors of original publications. The project website is at: `http://snap.stanford.edu/crank`. The website contains preprocessed data used in the paper and additional examples of CRANK's use. Source code of the CRANK method is available for download from the project website.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

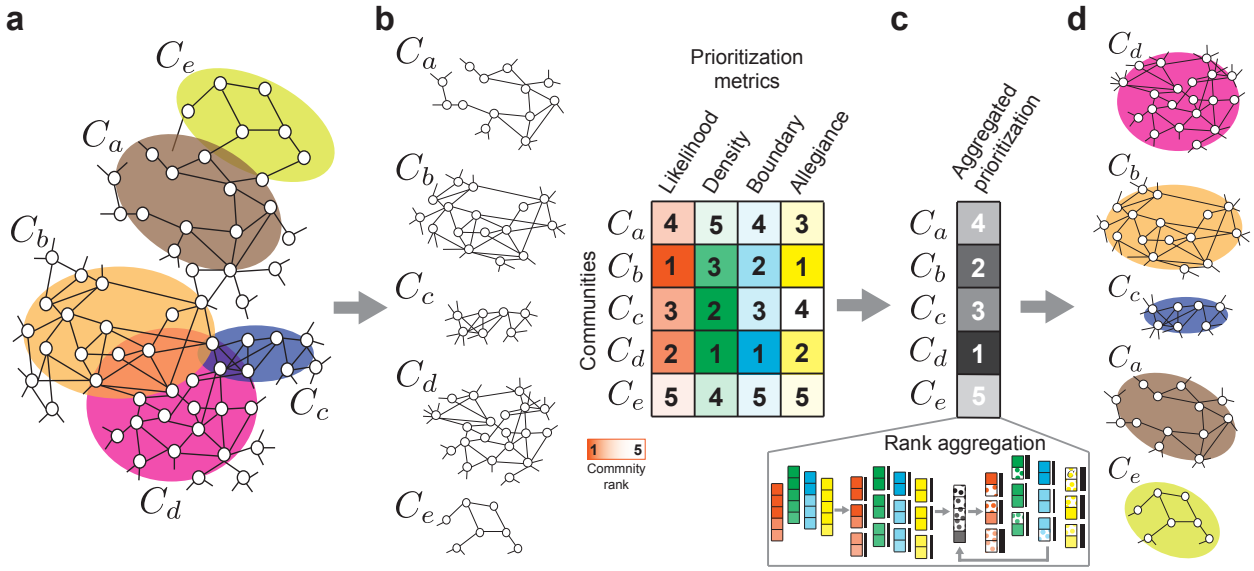

Figure 1: **Prioritizing network communities.** **(a)** Community detection methods take as input a network and output a grouping of nodes into communities. Highlighted are five communities, $(C_a, \ldots, C_e)$, that are detected in the illustrative network. **(b)** After communities are detected, the goal of community prioritization is to identify communities that are most promising targets for follow-up investigations. Promising targets are communities that are most associated with external network functions, such as cellular functions in protein-protein interaction networks, or cell types in cell-cell similarity networks. CRANK is a community prioritization approach that ranks the detected communities using only information captured by the network structure and does not require any external data about the nodes or edges of the network. However, when external information about communities is available, CRANK can make advantage of it to further improve performance (**SI Text** Sections S9 and S10). CRANK starts by evaluating four different structural features of each community: the overall likelihood of the edges in the community (Likelihood), internal connectivity (Density), external connectivity (Boundary), and relationship with the rest of the network (Allegiance). CRANK can also integrate any number of additional user-defined metrics into the prioritization without any further changes to the method. **(c)** CRANK then applies a rank aggregation method to combine the metrics and **(d)** produce the final ranking of communities.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

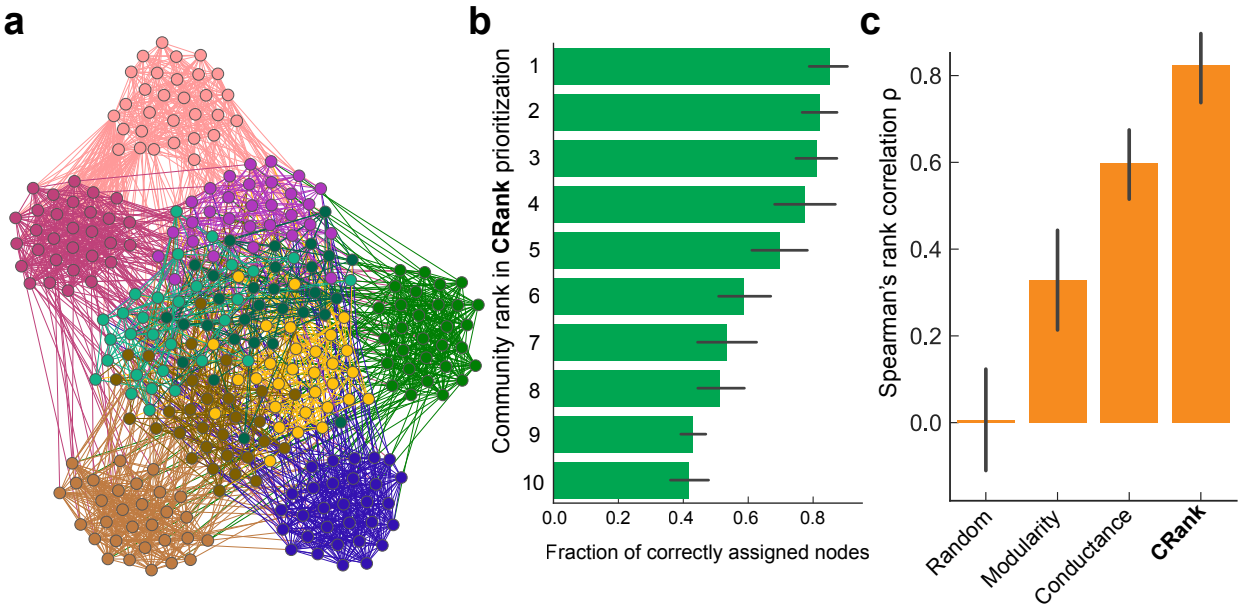

Figure 2: **Synthetic networks with planted community structure.** **(a)-(c)** In networks with known modular structure we can evaluate community prioritization by quantifying the correspondence between detected communities and the planted communities. **(a)** Benchmark networks on $N = 300$ nodes are created using a stochastic block model with 10 planted communities[10]. Each planted community has 30 nodes, which are colored by their planted community assignment. Planted communities use different values for within-community edge probability p_{in}, five use $p_{\text{in}} = 0.6$ and five use $p_{\text{in}} = 0.2$. As a result, planted communities with smaller within-community probability p_{in} are harder to detect. For each benchmark network we apply a community detection method[6] to detect communities and then use CRANK to prioritize them. CRANK produces a ranked list of detected communities. The gold standard rank of each community is determined by how accurately it corresponds to its planted counterpart. **(b)** Each bar represents one detected community and the bars are ordered by CRANK's ranking with the highest ranked community located at the top and the lowest ranked community located at the bottom. As a form of validation, the width of each bar corresponds to the fraction of nodes in a community that are correctly classified into a corresponding planted community, with error bars showing the 95% confidence intervals over 500 benchmark networks. A perfect prioritization ranks the bars by decreasing width. Notice that CRANK perfectly prioritizes the communities even though it only uses information about the network structure, and has no access to information about the planted communities. **(c)** Prioritization performance is measured using Spearman's rank correlation ρ between the generated ranking and the gold standard ranking of communities. A larger value of ρ indicates a better performance. Across all benchmark networks, CRANK achieved average Spearman's rank correlation of $\rho = 0.82$. Alternative approaches resulted in poorer average performance: ranking based on modularity and conductance achieved $\rho = 0.33$ and $\rho = 0.60$, respectively, whereas random prioritization obtained $\rho = 0.00$.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

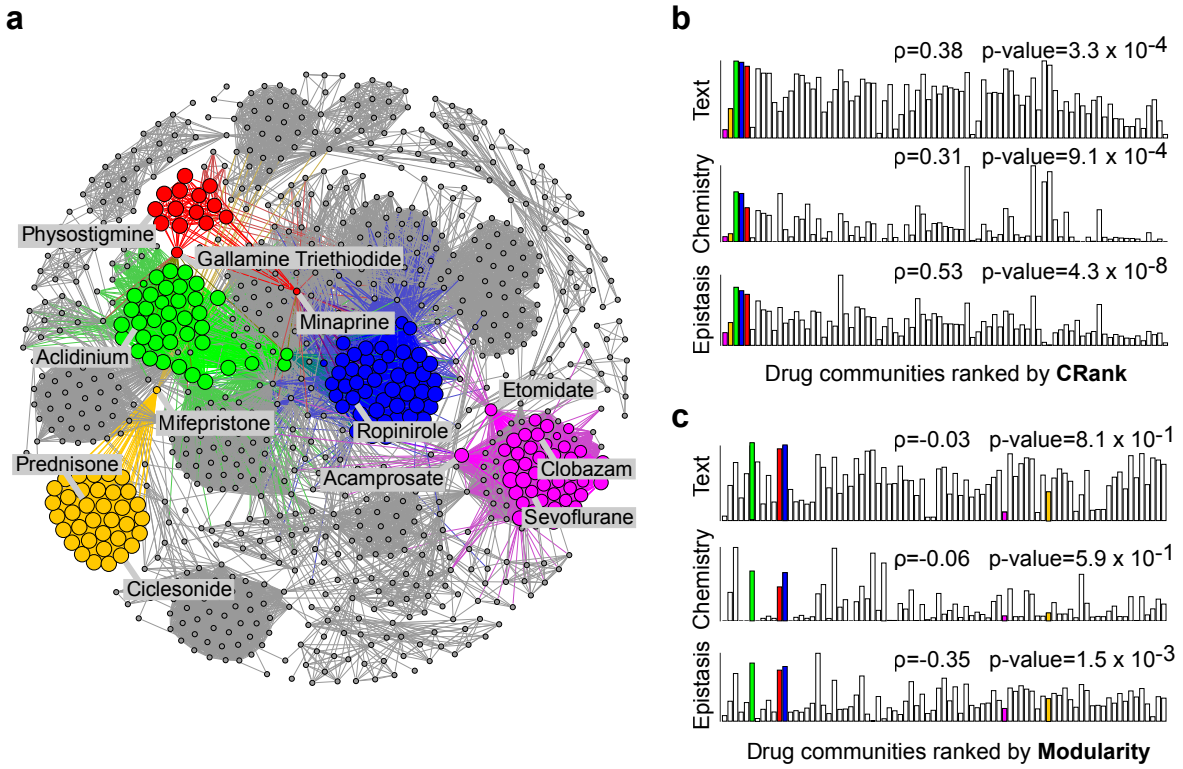

Figure 3: **Prioritizing network communities in the network of medical drugs. (a)** The network of medical drugs connects two drugs if they share at least one target protein. Communities were detected by a community detection method[7], and then prioritized by CRANK. Highlighted are five highest ranked communities as determined by CRANK. Nodes of the highlighted communities are sized by their score of the Likelihood prioritization metric (**SI Text** Section S3.1). Investigation reveals that these communities contain drugs used to: treat asthma and allergies (*e.g.*, prednisone, ciclesonide; yellow nodes), induce anaesthesia or sedation (*e.g.*, clobazam, etomidate, sevoflurane, acamprosate; magenta nodes), block neurotransmitters in central and peripheral nervous systems (*e.g.*, physostigmine, minaprine, gallamine triethiodide; red nodes), block the activity of muscarinic receptors (*e.g.*, acidinium; green nodes), and activate dopamine receptors (*e.g.*, ropinirole; blue nodes). **(b-c)** We evaluate community prioritization against three external chemical databases (**SI Text** Section S6.2) that were not used during community detection or prioritization. For each community we measure: (1) drug-drug interactions between the drugs ("Epistasis"), (2) chemical structure similarity of the drugs ("Chemistry"), and (3) associations between drugs derived from text data ("Text"). We expect that a true high-priority community will have more drug-drug interactions, higher similarity of chemical structure, and stronger textual associations between the drugs it contains. Taking this into consideration, the external chemical databases define three gold standard rankings of communities against which CRANK is evaluated. Bars represent communities; bar height denotes similarity of drugs in a community with regard to the gold standard based on external chemical databases. In a perfect prioritization, bars would be ordered such that the heights would decrease from left to right. **(b)** CRANK ranking of drug communities outperforms ranking by modularity **(c)** across all three chemical databases (as measured by Spearman's rank correlation ρ with the gold standard ranking). CRANK ranking achieves $\rho = 0.38, 0.31, 0.53$, while modularity obtains $\rho = -0.03, -0.06, -0.35$.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

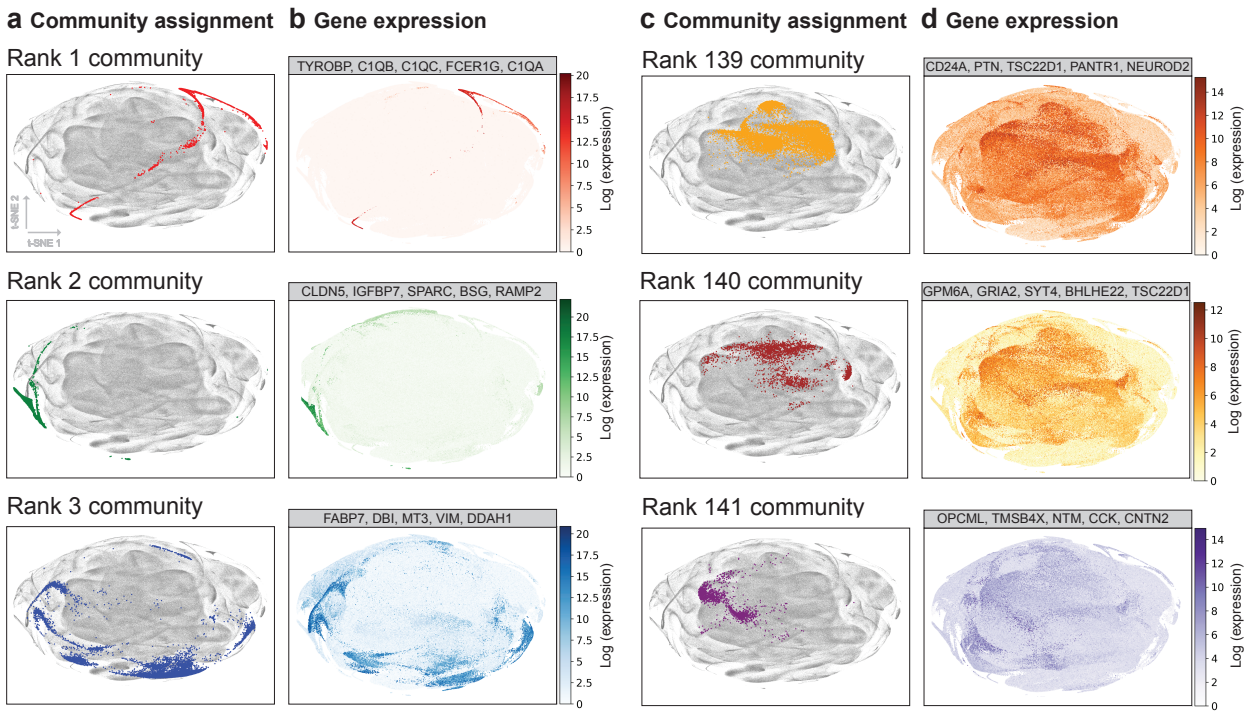

Figure 4: **Prioritizing network communities in the megascale cell-cell similarity network.** The network of embryonic mouse brain has 1,306,127 nodes representing brain cells[29]. Communities are detected using a community detection method developed for single-cell RNA-seq data[8] and prioritized using CRANK, generating a rank-ordered list of detected communities. **(a)** t-SNE projections[39] show cells assigned to 3 communities that are ranked high by CRANK (from top to bottom: rank 1, rank 2, and rank 3 community). t-SNE is a dimensionality reduction technique that is particularly well suited for visualization of high-dimensional data. Cells assigned to each community are distinguished by color, and all other cells are shown in grey. **(b)** We use the single-cell RNA-seq dataset to obtain a gene expression profile for each cell, indicating the activity of genes in the cell. For each community we then identify marker genes, *i.e.*, genes with the strongest differential expression between cells assigned to the community and all other cells[9]. In the t-SNE projection we then color the cells by how active the marker genes are. This investigation reveals that communities ranked high by CRANK are represented by clusters of cells whose marker genes have a highly localized expression. For example, marker genes for rank 1 community (the highest community in CRANK ranking) are *TYROBP*, *C1QB*, *C1QC*, *FCER1G* and *C1QA*. Expression of these genes is concentrated in cells that belong to the rank 1 community (top row). Similarly, marker genes for rank 2 and rank 3 communities are specifically active in cell populations that match well the boundary of each community (middle and bottom row). **(c)** t-SNE projections show cells assigned to 3 low-ranked communities (from top to bottom: rank 139, rank 140, and rank 141 community). **(d)** t-SNE projections are produced using the same differential analysis as in **(b)**. Although these communities correspond to clusters of cells (see t-SNE projections in **(c)**), their marker genes have diluted gene expression that is spread out over the entire network, indicating that CRANK has correctly considered those communities to be low priority.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

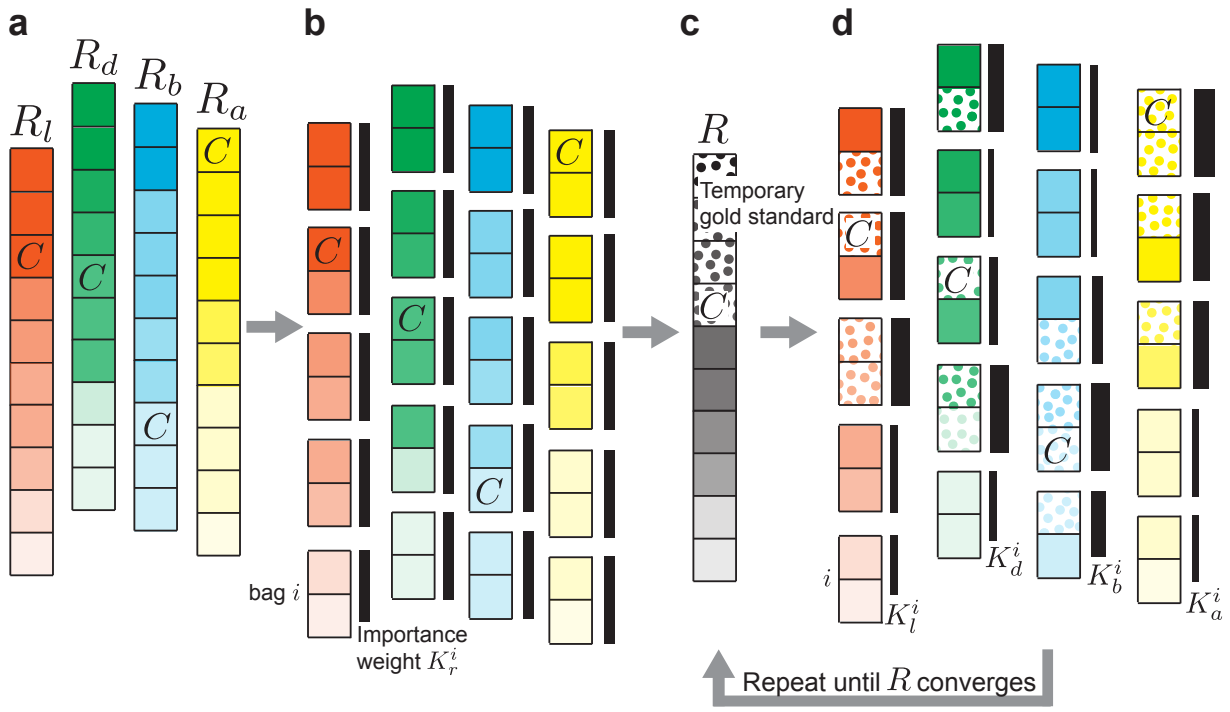

Figure 5: **Combining community prioritization metrics without an external gold standard. (a)** The rank aggregation algorithm starts with four ranked lists of communities, R_r, each one arising from the values of a different community prioritization metric r (where r is one of "l" – likelihood, "d" – density, "b" – boundary, "a" – allegiance). Communities are ordered by the decreasing value of the metric. We use C to indicate the rank of an illustrative community by the community prioritization metrics and at different stages of the algorithm. **(b)** Each ranked list is partitioned into equally sized groups, called bags. Each bag i in ranked list R_r has attached importance weight K_r^i whose initial values are all equal (represented by black bars all of same width). CRANK uses the importance weights K_r^i to initialize aggregate prioritization R as a weighted average of community ranks R_l, R_d, R_b, R_a. **(c)** The top ranked communities (denoted as dotted cells) in the aggregated prioritization R serve as a temporary gold standard, which is then used to iteratively update the importance weights K_r^i. **(d)** In each iteration, CRANK updates importance weights using the Bayes factor calculation[36] (**SI Text** Section S4). Given bag i and ranked list R_r, CRANK updates importance weight K_r^i, based on how many communities from the temporary gold standard appear in bag i. Updated importance weights then revise the aggregated prioritization in which the new rank $R(C)$ of community C is expressed as: $R(C) = \sum_r \log K_r^{i_r(C)} R_r(C)$, where $K_r^{i_r(C)}$ indicates the importance weight of bag $i_r(C)$ of community C for metric r, and $R_r(C)$ is the rank of C according to r. By using an iterative approach, CRANK allows for the importance of a metric not to be predetermined and to vary across communities.

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Author contributions. M.Z., R.S. and J.L. designed and performed research, contributed new analytic tools, analyzed data, and wrote the paper.

Author information. The authors declare no conflict of interest. Correspondence should be addressed to J.L. (jure@cs.stanford.edu).

Additional information. **SI Text** contains a detailed description of the community prioritization approach, descriptions of datasets, experimental setup, and additional experiments. **SI Table** contains detailed community prioritization results for the medical drug network. Code to run CRANK and examples are at: http://snap.stanford.edu/crank.

1. Benson, A. R., Gleich, D. F. & Leskovec, J. Higher-order organization of complex networks. *Science* **353**, 163–166 (2016).

2. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).

3. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. *Science* **353**, 1381 (2016).

4. Fortunato, S. Community detection in graphs. *Physics Reports* **486**, 75–174 (2010).

5. Newman, M. E. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**, 8577–8582 (2006).

6. Gopalan, P. K. & Blei, D. M. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences* **110**, 14534–14539 (2013).

7. Yang, J., McAuley, J. & Leskovec, J. Detecting cohesive and 2-mode communities indirected and undirected networks. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 323–332 (2014).

8. Levine, J. H. *et al.* Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).

9. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015).

10. Newman, M. & Clauset, A. Structure and inference in annotated networks. *Nature Communications* **7**, 11863 (2016).

11. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).

12. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

13. Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. EnrichNet: network-based gene set enrichment analysis. *Bioinformatics* **28**, 451–457 (2012).

14. Baryshnikova, A. Systematic functional annotation and visualization of biological networks. *Cell Systems* **2**, 412–421 (2016).

15. Hofree, M. *et al.* Network-based stratification of tumor mutations. *Nature Methods* **10**, 1108–1115 (2013).

16. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

17. Regev, A. *et al.* The Human Cell Atlas. *bioRxiv* 121202 (2017).

18. da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proceedings of the National Academy of Sciences* **112**, 12549–12550 (2015).

19. Layeghifard, M., Hwang, D. M. & Guttman, D. S. Disentangling interactions in the microbiome: a network perspective. *Trends in Microbiology* **25**, 217–228 (2017).

20. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nature Reviews Chemistry* **1**, s41570–017 (2017).

21. Schaeffer, S. E. Graph clustering. *Computer Science Review* **1**, 27–64 (2007).

22. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nature Communications* **7** (2016).

23. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Research* **42**, 1091–1097 (2014).

24. Erkinjuntti, T. *et al.* Efficacy of Galantamine in probable vascular dementia and Alzheimer's disease combined with cerebrovascular disease: a randomised trial. *The Lancet* **359**, 1283–1290 (2002).

25. Cecilia Rodrigues Simoes, M. *et al.* Donepezil: an important prototype to the design of new drug candidates for Alzheimer's disease. *Mini Reviews in Medicinal Chemistry* **14**, 2–19 (2014).

26. Kessenbrock, K. *et al.* Diverse regulation of mammary epithelial growth and branching morphogenesis through noncanonical Wnt signaling. *Proceedings of the National Academy of Sciences* **114**, 3121–3126 (2017).

27. Kellis, M. *et al.* Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences* **111**, 6131–6138 (2014).

28. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* **25**, 1491–1498 (2015).

29. Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications* **8**, 14049 (2017).

30. Haghverdi, L., Buettner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods* **13**, 845–848 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/315903; this version posted May 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

31. Zeisel, A. *et al.* Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).

32. Ma, J., Jiang, T., Tan, L. & Yu, J.-T. TYROBP in Alzheimer's disease. *Molecular Neurobiology* **51**, 820–826 (2015).

33. Agrawal, M., Zitnik, M. & Leskovec, J. Large-scale analysis of disease pathways in the human interactome. In *Pacific Symposium on Biocomputing*, vol. 23, 111 (World Scientific, 2018).

34. Choobdar, S. *et al.* Open community challenge reveals molecular network modules with key roles in diseases. *bioRxiv* 265553 (2018).

35. Karrer, B., Levina, E. & Newman, M. E. Robustness of community structure in networks. *Physical Review E* **77**, 046119 (2008).

36. Kass, R. E. & Raftery, A. E. Bayes factors. *Journal of the American Statistical Association* **90**, 773–795 (1995).

37. Berger, J. O. & Pericchi, L. R. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91**, 109–122 (1996).

38. Casella, G. & Moreno, E. Assessing robustness of intrinsic tests of independence in two-way contingency tables. *Journal of the American Statistical Association* **104**, 1261–1271 (2012).

39. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).