Lecture Notes in Social Networks

Mosab Alfaqeeh
David B. Skillicorn

# Finding Communities in Social Networks Using Graph Embeddings

Springer

# Lecture Notes in Social Networks

Lecture Notes in Social Networks (LNSN) comprises volumes covering the theory, foundations and applications of the new emerging multidisciplinary field of social networks analysis and mining. LNSN publishes peer- reviewed works (including monographs, edited works) in the analytical, technical as well as the organizational side of social computing, social networks, network sciences, graph theory, sociology, semantic web, web applications and analytics, information networks, theoretical physics, modeling, security, crisis and risk management, and other related disciplines. The volumes are guest-edited by experts in a specific domain. This series is indexed by DBLP. Springer and the Series Editors welcome book ideas from authors. Potential authors who wish to submit a book proposal should contact Annelies Kersbergen, Publishing Editor, Springer
e-mail: annelies.kersbergen@springer.com

Mosab Alfaqeeh • David B. Skillicorn

# Finding Communities in Social Networks Using Graph Embeddings

Springer

Mosab Alfaqeeh
School of Computing
Queen's University
Kingston, ON, Canada

David B. Skillicorn
School of Computing
Queen's University
Kingston, ON, Canada

ISSN 2190-5428 ISSN 2190-5436 (electronic)
Lecture Notes in Social Networks
ISBN 978-3-031-60915-2 ISBN 978-3-031-60916-9 (eBook)
https://doi.org/10.1007/978-3-031-60916-9

# Preface

Humans have always participated in communities. Historically, these communities were tribes and tribal divisions; today, communities are more likely to be groups of people who share kinship, culture, religion, interests, and jobs.

Identifying communities seems natural to us, but this is mostly because we fail to see how much semantics or understanding of human culture we use to do so. Given a group of 40 people, an outsider might be able to tell that it is an extended family because of physical resemblances, but would struggle to find the commonalities if it were a group of hobbyists or a work group. And even then, communities often adopt some kind of sign to indicate group membership: clothing, hairstyle, badges, accents, and so on.

When we come to try and identify communities algorithmically, few datasets have these rich markers of culture or even insignia, so the task is much more difficult. Although community detection has been a topic of research for decades, the rise of large-scale social networks has provided new motivation, as well as new difficulties because of their large scale. Some of the motivations for community detection in social networks are: to help individuals find others who are like them (so called *link prediction* or *friend recommendation*), or to find subgroups of individuals who can become the target of focused advertising or marketing. Most social network data contains little semantics—perhaps only a single type of relationship between pairs of individuals—so that finding useful communities is extraordinarily difficult.

Community detection techniques have been studied in several different research communities, and have been published in different venues. Because of the specific formulation of the problem to be addressed, these techniques are not easy to compare. In particular, performance comparisons across the board are rare—each paper describes performance on a few datasets that differ from those used in other papers. It is difficult to discern the state of the art because of this variability.

The key to finding communities is to define a notion of *similarity*—what is it that makes two individuals similar and, usually, how similar. This may be derived from explicit markers—the two individuals have chosen to be related—or from some similarity of the content that they post, or from their activities visible in the social network. Recent work has tried to leverage both structural similarity and content

similarity, raising the immediate problem of how to weight these two different properties relative to one another. It turns out that content similarity in different modalities, that is different ways of expressing content, is critical and we pursue this, showing that not doing so leads to much poorer performance.

We develop a technique for finding communities that begins from several different kinds of similarity, expressed as weighted graphs, combines these graphs in a principled way, and then embeds them in a geometry. Finding communities in a geometry reduces to clustering, for which conventional techniques can be used. We show that this technique outperforms a wide range of known community detection techniques on data crawled from Instagram, using a range of commonly used performance metrics. We also show that the communities we find have much better content coherence, showing that their members really do share common interests. In contrast, many of the other techniques, although producing communities that score well on popular metrics, do not have content coherence. This suggests that some commonly used community detection techniques contain subtle flaws.

We also consider the role of directed relationships between pairs of users, and show that these tend to break communities. Directed edges are natural in some social settings, and this suggests that the problem of community detection needs to be rethought in such networks.

This book builds on the doctoral work of the first author.

Kingston, ON, Canada
April 2024

Mosab Alfaqeeh
David B. Skillicorn

# Contents

# Introduction


**Abstract** Communities represent groups of objects, usually individuals, who are similar to one another, and distinct from those in other communities. This intuitive description is difficult to make rigorous, and has led to different threads of research that are unknown to each other. Community detection techniques often start by defining what it means to be similar, and then developing algorithms to find sets of similar objects. Members of a community can be similar because they are connected to one another explicitly by a declared relationship, or they may be similar because they have shared interests. However, better communities are found when both kinds of similarity are used together. This means deciding how to balance the two modalities to produce the best overall similarity measurement.

**Keywords** Natural communities · Similarity · Structural similarity · Attribute similarity · Social networks

## 1 Communities Are Natural But Subtle

For millennia humans have lived in communities, in tribal groups and, within these, in families. We can immediately see some of the complexities in looking within social groups for communities. Tribes are relatively well defined, by geography and culture, and their borders tend to be closed. Moving from one tribe to another involved rituals (marriage) or war (prisoner capture) as well as physical movement.

Defining a community comprising a family is a more difficult task, for awareness of obligations, and for knowing properties like who must be invited to a wedding. The variety of ways in which families can be defined and delineated illustrates some of the complexities that also apply to other kinds of communities.

Communities are important for constraining interactions in a world of finite resources. Nobody can interact with everyone, so people form communities with which they share common interests, and businesses try to categorise their customers into communities that can be served differently, in ways customised to their needs. *Dunbar's number* [8], estimated to be about 150, represents the number of others



M. Alfaqeeh, D. B. Skillicorn, *Finding Communities in Social Networks Using Graph Embeddings*, Lecture Notes in Social Networks,
https://doi.org/10.1007/978-3-031-60916-9_1

with whom any individual can maintain a connection that could be plausibly be strong enough to be called a relationship. It was first posited for tribal societies but seems to carry over well to the online world, suggesting that it is based in the "hardware" of human social cognition. These 150 people also seem to exist in a structured form called the "Rule of Almost Threes": people tend to have (slightly more than) 3 close connections, usually immediate family: $3^2 = 9$ quite close connections; $3^3 = 27$ regular connections; and $3^4 = 81$ acquaintances, a total that is close to Dunbar's number. Thus we expect that communities that are maintained by active social interactions will not easily exceed this size. On the other hand, communities that exist because of weaker properties, say, shared interests or common buying patterns can be much larger.

It is also true that individuals often belong to more than one community simultaneously. Tribes and families tend to be absolute divisions—everyone is in one tribe and (mostly) in one family—but a individual can belong to many different communities based on interests and activities: work groups, sports teams, gyms, cultural groups, hobby groups, those who love to travel or to eat in restaurants, and many more. Sorting a group of people into communities must rely on the fact that relationships are of different kinds. A family comprises people with a genetic and/or regulated formal connection; a nationality comprises people with a common birthplace or a formal induction ceremony, a work group comprises people who work together, a hobby group comprises people who share the same interest in a non-work activity. Trying to find communities without paying attention to different kinds of relationships is doomed to failure.

And some kinds of communities do not have explicit relationships between their members. Rather they share an interest, perhaps gourmet Thai food. This ties them together, so a Thai restaurant chain may want to identify them, but they may not even know one another. So a community like this is quite a different thing from a tribe.

Communities are complex objects and, as we shall see, there has been considerable disagreement about how to frame them, how to find them, and how to assess how well techniques to find them perform.[1]

## 2 What Is a Community?

We now begin to define, more carefully, the problem of community detection. Communities represent groups, usually of individuals, who are internally similar to one another and, by implication, different in some way from those in other communities. This includes the examples we have just discussed: sometimes groups who are connected to one another, and sometimes groups whose connections are

[1] There is a game called *Connections* https://www.nytimes.com/games/connections where the aim is to put words into groups. The number of different criteria that can be used to do this is a vivid illustration of the complexity of the problem.

implicit. At their simplest, communities are disjoint, that is each individual is a member of exactly one community; but is possible, and perhaps more realistic, to consider individuals are being members of more than one community.

Social networks are an environment where communities are certainly present: families keeping touch with one another, work teams, leisure activity groups, and collections of people with a common interest. It differs from more traditional environments in which there are communities because of scale—the number of participants is typically in a hundreds of millions—and the ease with which a "connection" can be made. In a conventional (real-world) social network a connection uses up one of the 150 slots implied by Dunbar's number. This is no longer true in the online world, creating the impression that people really can have thousands of "friends". One of the challenges, then, is to separate real connections from illusory connections.

Social networks provide vast amounts of data about individuals, who describe themselves explicitly by facts about themselves that they reveal in a "bio" and who they choose to make connections with; and implicitly by the content that they choose to post. There are many practical reasons why it is interesting and useful to find communities in social networks: it can improve 'friend' recommendations, it can provide more targeted content, advertisements or newsfeeds, and it can help to find individuals of particular interest, whether influencers or terrorists [3, 4].

The intuitive idea of a community that we have been discussing turns out to be surprisingly difficult to make rigorous. Even the terminology has been confusing: communities are sometimes called clusters, modules, or subgraphs [5, 6, 10].

The key idea that lies behind communities is *similarity*. Each pair of individuals in a community must be similar enough to one another, and all the individuals in one community must be dissimilar enough to all of those in other communities. Part of the complexity of community detection is that there are many plausible, even useful, ways to define similarity. And of course there is considerable room to choose the threshold to be similar *enough*.

In social networks there is typically a way for each user to declare explicitly that someone else is similar to them by creating a link to them. This link goes by a number of names ('friend', 'follower', 'connection') but all are created as the result of actions. This usually requires an action by both individuals at the ends of the potential link: a request from one and an acceptance by the other,

Creating similarity links produces a graph whose nodes are users and whose edges represent these explicit connections. The edges may be typed (say, acquaintance vs relative) but this is hardly ever done (although several social media platforms use this data internally). Edges may also be weighted by the intensity of the connection, which could be declared by the users or set to some default weight and adjusted up or down based on the interactions that happen between them.

Another way in which users can be similar is because they have similar properties and interests, expressed in what they say about themselves and post about their lives. These attributes may be revealed explicitly in a *bio* where they reveal demographic data about themselves, for example, gender, location, age, and interests. The number of such attributes is usually fixed and the number of values that an attribute can have is usually limited as well.

Users' attributes can also be revealed implicitly by the content of what they post, most often natural language or images. These implicit attributes are naturally open ended [13] which means that they can be more revealing but also more difficult to work with.

These two kinds of similarity—structural similarity, based on explicit connections—and attribute similarity, based on similarity of interests, need not agree. It is possible, for example, for two individuals to have very closely aligned interests but never to have encountered one another, and so never had the opportunity to connect explicitly.

Figure 1 shows a small example of a social network graph where each vertex represents a user's profile and the edges represent the explicit connections between pairs of users. In addition, each user has an interest, expressed as an attribute of its node.

If we want to find communities in this social network, there are several different plausible answers.

Figure 2 is a clustering based on structure, the explicit links between the user pairs. Users in the same cluster belong together from a relationship perspective, but they do not all have the same interests.

Figure 3 is a clustering based on similarity of attribute values, that is, the interests associated with each profile. Users in the same cluster belong together because they share interests, but they do not necessarily have relationships with the others in their cluster.

Figure 4 is a clustering that is based on both structure and attribute information. Users within each cluster are mostly connected, but they are also mostly homogeneous in interests. Such a clustering seems intuitively to be a better representation of the actual similarities between groups of users within the social network.

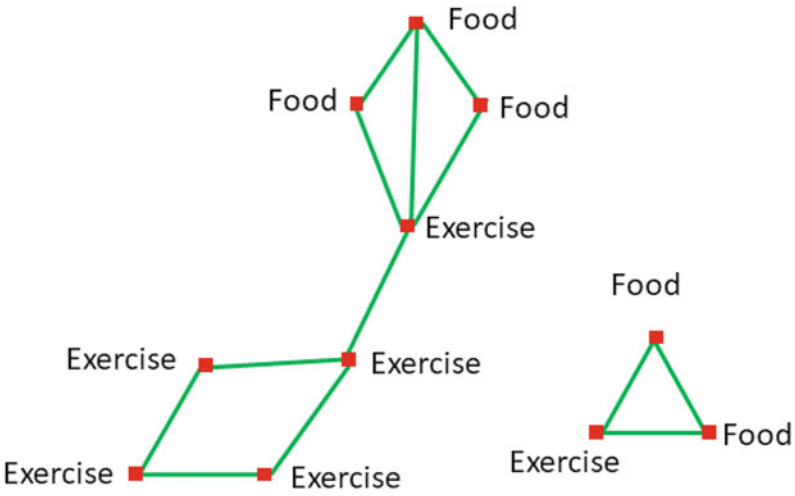


**Fig. 1** Graph of explicit connections and implicit similarity of interests

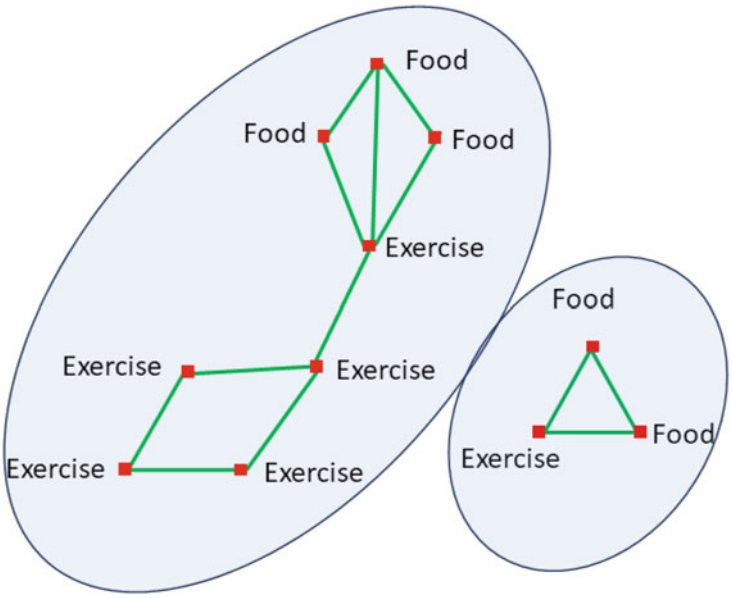


**Fig. 2** Structure-based clustering of the graph

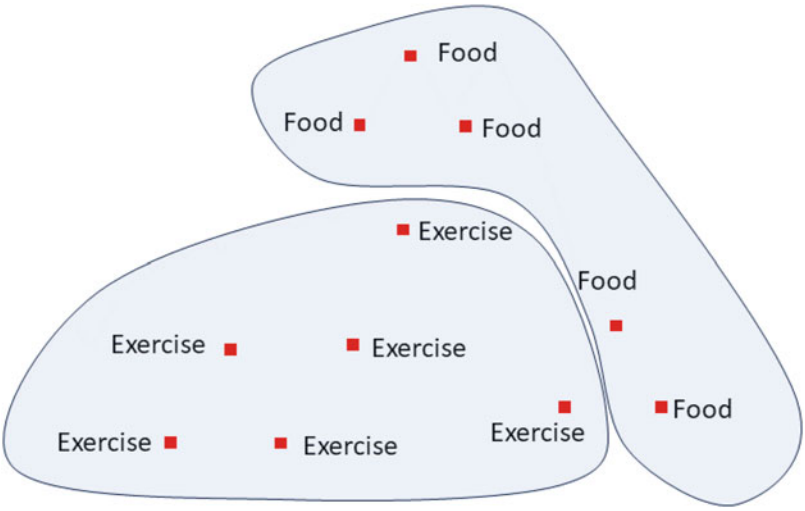


**Fig. 3** Attribute-based clustering of the graph

Most social network platforms allow for the values of a fixed set of attributes to be set by each user and so, of course, there are only a limited number of ways to be similar. Even when attributes are available, some users may not bother to specify them, especially if they are complex, or may be concerned about their privacy. In

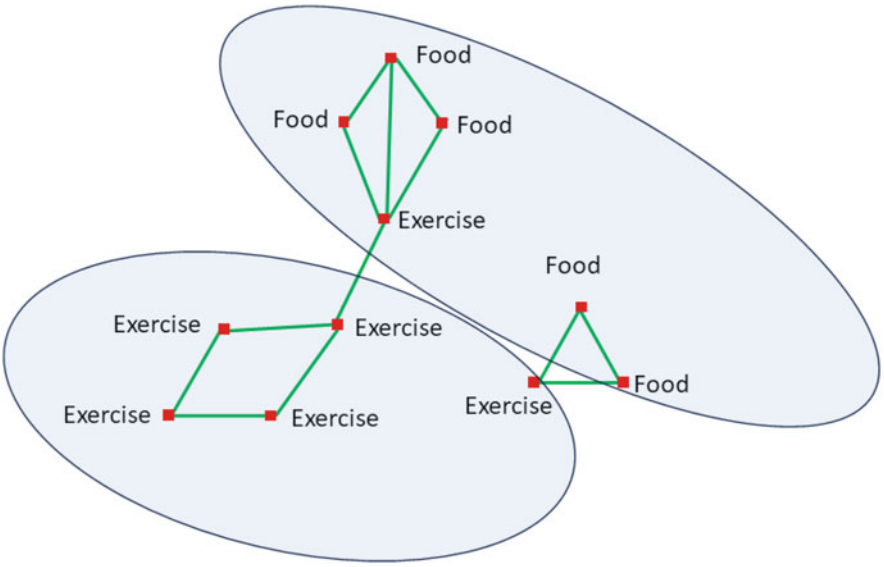


**Fig. 4** Structural and attribute clustering

one dataset [9] containing 19,624 users, nearly 90% of user gender information had been specified, but only 12.36% of users' birthdays, and 22.48% of their relationship status.

Other attribute values can be inferred from the content that users post, including some attributes that might, in other settings, have been specified explicitly. Attributes such as age, gender, regional background, location, political views, and sentiment or attitudes [1, 15, 16, 18, 23] can be inferred from natural language with reasonable accuracy.

In ordinary interactions, humans speak about themselves between 30 and 40% of the time, but this increases to around 80% online [12], and online presentation allows for careful construction of the persona being presented. However, much natural language has a strong subconscious component and so is outside the explicit control of authors. Algorithmic analysis often sees aspects of users that they are unaware that they are revealing.

There are advantages and disadvantages to inferring user properties from the content of whatever they post. On the one hand, social and cultural processes encourage revealing personal data to an extent that would have been unthinkable even a few decades ago. On the other hand, there tends to be a bias towards revealing only positive attributes, and perhaps also to be misleading or even deceptive. It is well known that users tend to post primarily the positive aspects of their lives, creating unrealistic expectations in those who see their profiles.

The open-ended content that users can post on social network platforms is of several qualitatively different kinds, although the details differ from platform

to platform. Commonly, users can post text, both descriptive of themselves and commentary about what's happening in their lives and the world; images that describe their experiences or that resonate with them, perhaps photographs that they have taken; and hashtags, a form of text with special properties. Some social network platforms focus on textual content, and others on visual content.

## 3 Motivation

There are several reasons why finding communities in online social network platforms can be useful:

- Social network companies support themselves by serving advertisements to their users. These can be personalised, allegedly targeting products and services that are of interest to each individual user. This requires considerable expense in (paying for) collecting the required data and modelling each user, and its success remains questionable. Also increasingly there is public pressure to allow opting out of some aspects of this personalisation process.
  At the other extreme, advertisements can be generic, so that all users in certain broad groups (location of connection, time of day, known demographics) see the same ones,
  Clustering users creates a new possibility: serving advertisements to the members of a community targeted to the interests of that community. This middle ground has advantages over both of the extremes. Communities only need to be found once and finding them is totally under the control of the platform. Once found, the platform can be assured that the advertisements are fairly well targeted.
- On many platforms, users want to find more content similar to something they have already seen. Search may partially solve this problem but some search terms are too generic or ambiguous to find all relevant content. In a recent test, searching for "apple brand" at Google produced pages about the tech company but searching for "apple brands" produced pages about the fruit. And that is from the leader in online search.
  A platform with knowledge of its communities can restrict search to within the community of the user doing the searching, with greater likelihood of finding relevant and satisfying content.
  Platforms that proactively feed new content to users can use community information to choose content that users will find relevant because it fits with the interests of their community.
- Another way in which platforms maintain user interest (and hence exposure to advertisements) is by making recommendations, not of new content, but of new people. In other words, platforms suggest, to a user, other users that are not already known but 'should be'. One obvious way to find such people is to look at the content of both users' profiles and notice that it is similar. Since they

already have the same interests, they may want to interact with one another more explicitly.

More subtly, two users can be similar because they see the same graph landscape from the perspective of the graph of explicit links. In the simplest case, this could just be that they have (almost) the same set of friends (but in this case they almost certainly already know one another). But the same property can hold at greater scales, and reflects a deeper, though less obvious, form of similarity.

Making "new connection" ("friend") recommendations within a community increases the likelihood that the two users involved really are similar and so are likely to connect.

- Since advertisements themselves have proven limited, marketers have turned to *influencers*, users who either have a wide reach within a social network, perhaps by having many connections themselves, or having a track record of *virality*, having their content read and copied by many others so that it spreads through the social network like a virus.

  The problem for marketers is how to find a selection of influencers who can, together, reach the desired members of the social network. Communities can help with this problem too. A community is a natural way of describing a target audience. And influencers can be chosen so that there is one influencer for each target community. In other words, communities help with the coverage problem from both ends.
- Social network platforms are plagued with *fake accounts* or *bots* which are created to influence via posting and disseminating information, often false. They have been used to influence election outcomes, to manipulate stock markets, to spread links to websites, driving traffic to them, and to provide credibility for users or organisations that can then be used for criminal purposes such as cyber attacks. The fraction of social media accounts that are fake has been estimated to be between 5% and 15% of all accounts. Their presence therefore distorts all measurements of social network characteristics, and much ordinary user activity on a platform.

  It is, however, difficult to create an account that is convincingly human because it requires consistency among the various attributes in both the bio and the posted content, and it must be plausibly rich with images and hashtags. It is difficult to automatically create consistent content when it is expressed in qualitatively different modalities, especially at scale. If the post content is about, say, travel then which and how many travel-related images should be included to support a supposed interest in travel without appearing like a random selection of place images? And where are these images to be obtained from when free image sites are off limits because they are easy to check? If the expression of content via different modalities is explicitly assessed then it is much more expensive to create plausible fake accounts.

  Fake accounts also create connections in an attempt to make particular user accounts seem more important or impactful than they actually are, and this creates patterns of connection that are different from those created by genuine users who typically add connections a few at a time.