The Prediction of Venture Capital Co-Investment Based on Structural Balance Theory

Zhiyuan Wang, *Member, IEEE*, Yun Zhou, Jie Tang, *Senior Member, IEEE*, and Jar-Der Luo

Abstract—Venture capital (VC) is of great importance to high-tech industry and network economy since many high-tech firms benefit from VC, especially when they are in their infancy, such as Google, PayPal, and Alibaba. Over 80 percent of the VC investments are related to at least two investors and so co-investment is an important phenomenon in the VC market. However, it is challenging to predict future co-investments due to the complexity and uncertainty of VC behavior. In this paper, we formulate the problem of co-investment prediction into a factor graph model incorporating structural balance theory. We design a large number of features from the perspective of both domain knowledge and social network, and select prominent features by group Lasso. In this paper, we introduce two new investment datasets for the study of VC. Experiment results demonstrate that the proposed model significantly (+9% in terms of accuracy) outperforms the baseline methods. It is shown that only the top 10 features selected by group Lasso (e.g., nationality, number of common neighbors, betweenness, shortest distance, investor type, number of invested fields, and Jaccard similarity of invested fields) can explain the formation of the VC network quite well (around 90 percent in terms of accuracy). In addition, we have some interesting findings. For instance, in the VC network, the co-investor of my co-investor tends to be my co-investor; VC pairs from the same country, of the same investor type, with short distance, with more common neighbors or with appropriate Jaccard similarity of invested fields are likely to co-invest; VCs of large betweenness or of a large number of invested fields have advantage in the VC network; investors of Asian countries, especially of China, are more likely to have social relations than other countries.

Index Terms—Venture capital, co-investment, prediction, factor graph model, group lasso

1 Introduction

Venture capital (VC) is financial capital provided to early-stage, high-potential, growth startup firms. VCs are unsung heroes behind high-tech firms, such as Google, PayPal and Alibaba, especially when the firms are in their infancy. Without VC, high-tech startups will suffer from a shortage of funds and business directions, and so more and more importance has been given to VC in the era of information technology and network economy.

High-tech industry routinely acknowledges that communities knit together by networks of social relations are essential for the development of the industry, and emphasizes that VCs hold central positions in these networks [1]. Based on the statistics on the free online CrunchBase dataset (cf. Section 3), 80.9 percent of VC investments are related to at least two investors, thus co-investment is an important phenomenon in the VC market. We cannot fully understand VC behavior without a detailed exploration of co-investment.

In this paper, we study the problem of predicting whether two VCs will co-invest or not in the near future, given the existing VC network. This research is also of great interest to practitioners in the field of investment. For instance, the study can help a VC manager to find co-investors from a large number of candidates automatically. However, due to complexity and uncertainty of VC behavior, it's challenging to accurately predict future co-investments, and we address the challenges as follows. First, what factors influence the formation of co-investment relationships? Second, how to select a small number of fundamental factors that best explain the formation without significant drop in performance? Third, how to design a mechanism that incorporates social network theory affecting the formation of co-investment relationships?

Co-investment has been studied for many years in sociology and economics, such as [1], [2], [3], [4], [5]. Lerner [2] studied the principle of who will be a good co-investor and when to reconstruct a co-investment. Sorenson and Stuart [3] studied the effect of geographic spaces on co-investment. Based on 45 years of VC data from the US, Kogut et al. [4] found several features that might have influence on new co-investments. Powell et al. [5] studied four kinds of effects on interorganizational collaboration. However, most of researches dealt with a small dataset with at most hundreds of VCs except [4], and they only explored a few features for co-investment without detailed analysis of contribution of different kinds of features. In addition, few works predicted future co-investment with a unified model and presented the performance of prediction.

Solution and contribution. In this paper, we formulate the problem of co-investment prediction in the VC network and perform a series of observations of the data. Based on the observations and structural balance theory, we propose a structural balanced factor graph model named SBFG to predict the co-investment at time $t+1$, given co-investment

- *Z. Wang and Y. Zhou are with the State Key Laboratory of High Performance Computing, National University of Defense Technology & School of Computer, National University of Defense Technology, Changsha 410073, China. E-mail: {wzy, zyvxd}@nudt.edu.cn.*
- *J. Tang is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: jietang@tsinghua.edu.cn.*
- *J.-D. Luo is with the Social Network Research Center, Tsinghua University, Beijing 100084, China. E-mail: jdluo@mail.tsinghua.edu.cn.*

Manuscript received 25 Aug. 2014; revised 14 June 2015; accepted 24 Aug. 2015. Date of publication 9 Sept. 2015; date of current version 6 Jan. 2016. Recommended for acceptance by A. Banerjee.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2477304

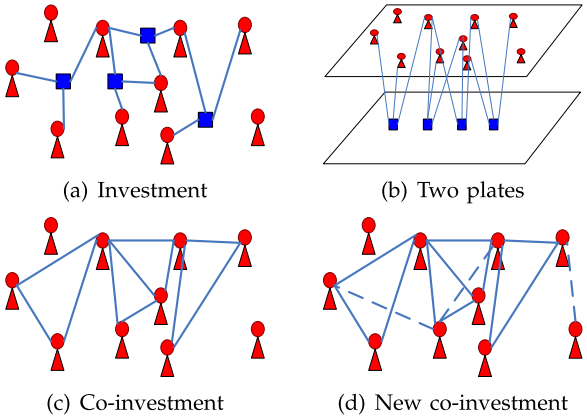


Fig. 1. Investment and co-investment. The red person represents a VC, and the blue box represents a startup. The line in (a) and (b) indicates an investment. The solid (dashed) line in (c) and (d) indicates the past (future) co-investment between VCs.

network of time span $\{1, t\}$. We develop an approximate algorithm using loopy belief propagation (LBP) to efficiently learn the proposed model. Experiment results demonstrate that the proposed model SBFG significantly (+9 percent in terms of accuracy) outperforms the baselines, i.e., logistic regression and SVM.

We design a larger number of features from the perspective of both domain knowledge and social network, which cover most features that have been proposed in past literature, such as [1], [2], [3], [4], [5]. In order to gain both interpretability of features and high accuracy, we select prominent features by group Lasso. It is shown that only the top 10 features selected by group Lasso can explain the formation of the VC network quite well (it drops by only 0.18 percent in terms of accuracy compared with a total of 81 features), e.g. nationality, number of common neighbors, betweenness, shortest distance, investor type, number of invested fields and Jaccard index of invested fields. We have some interesting findings by exploring the prominent features, which can be used to explain investment behavior in the VC market.

We introduce two new investment datasets for the academic community, which can be applied to the study of data mining or social network analysis.

Organization. Section 2 formulates the problem. Section 3 introduces the dataset. Section 4 describes feature design and feature selection by group Lasso. Section 5 presents the observation of the prominent features selected by group Lasso. Section 6 proposes the structural balance based factor graph (SBFG) model and learning algorithm. Section 7 presents the experiment results and detailed analysis. Section 8 further explores another dataset that focuses on the startups in China. Section 9 reviews the related work and Section 10 concludes the paper.

2 Problem Formulation

In this section, we first give an illustration of investment and co-investment, present the formal definition of co-investment and then propose a formal description of the problem. We formulate the problem in the context of VC to keep things concrete.

TABLE 1
Information and Statistics of CRUNCH

Item	CRUNCH (1984-2014)
Investment information	VC, Startup, Funded year, Round, Rasied amount
VC information	Investor type, Location, Field
Startup information	Field, Location
#Investment	90,282
#VC	18,716
#Startup	25,327

Fig. 1 shows the investment and co-investment in the capital market. In Fig. 1a, the red person represents a VC, the blue box represents a startup that gets funded, and a line between a VC and a startup represents an investment. VCs and startups are in heterogeneous spaces, which are denoted by two plates in Fig. 1b. To simplify the network, we consider the co-investment of VCs by adding a link between two VCs that invest in a common startup in the same year, as shown in Fig. 1c. Given the VC network in the past, we'd like to predict whether two VCs will co-invest or not in the near future, and the dashed lines in Fig. 1d are new co-investments in the near future.

Definition 1 (Co-investment). We say that two VCs co-invest in a given year, if they invest in the same startup(s) in the year. Accordingly, they call each other co-investor.

The number of investments increases over time, and the VC network $G^t = (V^t, E^t)$ are also evolving, where V^t is the set of accumulated VCs ($|V^t| = N$), and $E^t \subseteq V^t \times V^t$ is the set of accumulated co-investment relationships between VCs until time t. We are concerned with the following problem.

Problem 1. Predict whether two VCs will co-invest or not in the next year. Let $G^t = (V^t, E^t)$ be the VC network in time span $\{1, t\}$, given two VCs, the task is to predict whether they will co-invest or not in time $t + 1$.

It bears pointing out that our problem is quite different from existing link prediction problems [6], [7], [8]. First, the VC network is intrinsically dynamic and multi-dimensional, which are not well treated in the traditional link prediction research. Second, it is not clear what are the fundamental factors that influence the formation of the VC network. Finally, one needs to incorporate the different factors (e.g. social theories, statistics and our intuition) into a unified model to better explain the co-investment relationship.

3 Data Description

The dataset (CRUNCH) comes from the free online CrunchBase,[1] which is updated frequently. The dataset contains open investment events in the world from 1984 to 2014, and there are a total of 18,716 VCs, 25,327 startups, 90,280 investments and 152,227 co-investments. The original information and statistics of CRUNCH are summarized in Table 1.

In recent years, the VC investment developed very quickly. The distribution of investment over year is shown in Fig. 2. In the first fifteen years (1984-1998), the number of investments every year was less than 50, and it rose to 168

1. http://www.crunchbase.com/, March 20th, 2014

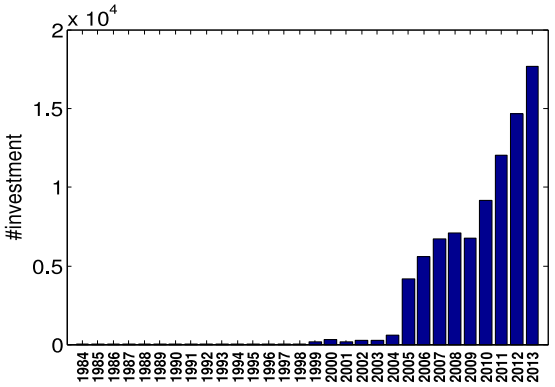
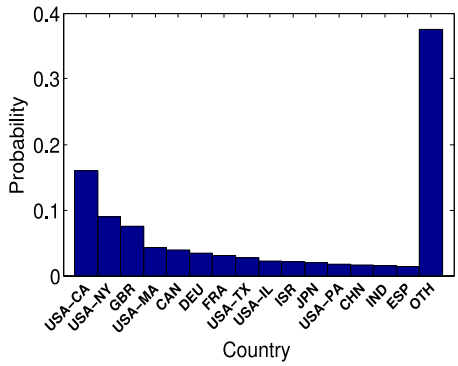
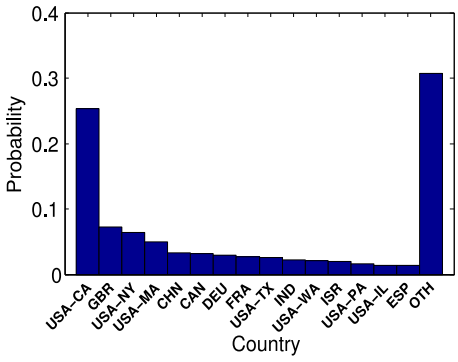


Fig. 2. Distribution of investment over year. The number of investments increases rapidly from 2005.



(a) VC distribution



(b) Startup distribution

Fig. 3. VC/startup distribution over country. The state of California is the most active area of VC and startup in the world.

in 1999. After six years of steady increase (1999-2004), the number jumped to 4,196 in 2005. From then on (2005-now), the number increases rapidly, with the exception of 2009, due to the economic depression.

The distribution of VC and startup over country/area are shown in Fig. 3, where the notation of three-letter denotes the country, which is defined in ISO 3166-1.[2] For instance, *USA* denotes the US, and *GBR* denotes United Kingdom. *OTH* denotes all other countries as a whole. Since VCs in the US account for more than half of total VCs in the world, the state of the US is treated as an entity in the statistics, where *CA* denotes California, *NY* denotes New York, and so on.

2. http://en.wikipedia.org/wiki/ISO_3166-1_alpha-3

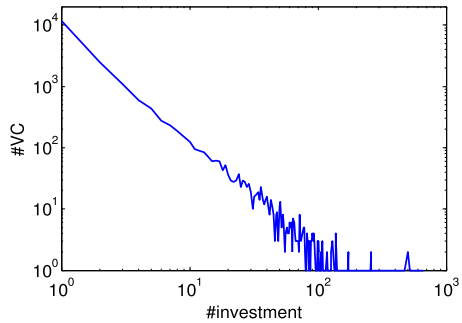


Fig. 4. Power-law distribution of VC over the number of investments (both *Y*-axis and *X*-axis have logarithmic scale). The investments of 7.6 percent VC firms account for 64.7 percent of all investments.

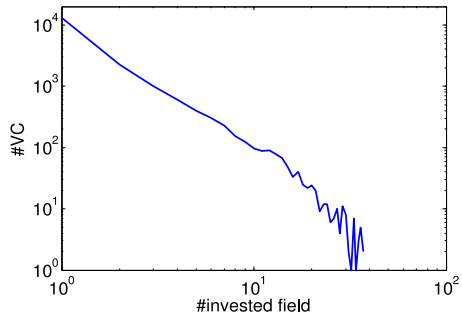


Fig. 5. Power-law distribution of VC over the number of invested fields (both *Y*-axis and *X*-axis have logarithmic scale).

Note that the number of VCs in the state of California is even larger than that of any other country in the world, and this area dominates the VC market of the world. Fig. 3a lists the top 15 countries/areas with the most VCs, and Fig. 3b lists the top 15 countries/areas with the most startups.

The difference between the numbers of investments of different VCs is very large. The distribution of VC over the number of investments is shown in Fig. 4, where the curve roughly obeys the power-law. From 1984 to 2014, every VC has 4.8 investments on average. The five VCs with the most investments are Sequoia Capital (659 investments), Start-Up Chile (607), Intel Capital (571), New Enterprise Associates (536) and Y Combinator (533). 62.7 percent of VCs have only one investment. The VCs with the more than 10 investments account for 7.6 percent of all VCs, and they have 64.7 percent of all investments.

In usual, VCs invest in several different fields to avoid risks. The distribution of VC over the number of invested fields is shown in Fig. 5, where the curve also roughly obeys power-law. There are 44 fields in CRUNCH (cf. Fig. 7). From 1984 to 2014, every VC invested in 2.2 fields on average. The VCs with the most invested fields are SV Angel, Start-Up Chile (37 invested fields), Kleiner Perkins Caufield & Byers, Sequoia Capital, Y Combinator, New Enterprise Associates and Techstars (36 invested fields).

The distribution of startup over the number of investments is shown in Fig. 6. Note that, different from Fig. 4 and Fig. 5, in Fig. 6, only the *Y*-axis has a logarithmic scale. From 1984 to 2014, every startup gets 3.6 investments on average. The five startups with the most investments are Fab (59 investments), ecomom (58), CardioDx (54), Practice Fusion (53) and Aperto Networks (49). 68.8 percent of startups are with less than or equal to 10 investments.

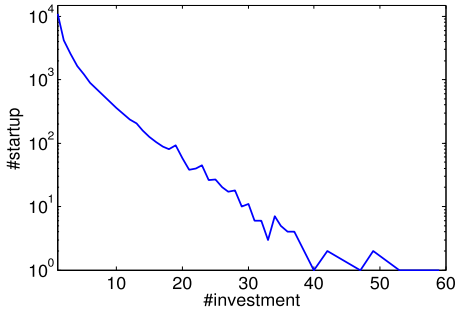


Fig. 6. Distribution of startup over the number of investments (*Y*-axis has a logarithmic scale).

The distribution of startup over the field is shown in Fig. 7. There are 44 fields in CRUNCH, where the top five fields are software (13 percent), biotech (8.9 percent), mobile (7.4 percent), web (7.3 percent) and enterprise (7.2 percent).

4 Feature Design and Selection

4.1 Feature Design

We design a large number of features for co-investment from the perspective of both domain knowledge and social network. According to the time characteristics, the features can be categorized into static features and dynamic features. The static feature does not change over time, such as nationality and investor type, but the dynamic feature changes over time, such as invested fields and betweenness on the VC network. Note that, the static feature takes the same value for different years, while the dynamic pattern should be normalized within the year (cf. Section 5.2), otherwise the values are not comparable between years. The dynamic features can be further divided into dynamic domain features and dynamic topology features. The former is related to domain knowledge of VC, and the latter is related to the evolving VC network.

All features for co-investment are summarized in Table 2. The features are self-explained, and the fifth column shows a short description of the features. In Section 5, we shall explain the important features in details.

4.2 Feature Selection with Group Lasso

As stated in the previous section, a large number of features have been taken into consideration. However, the relations among features are interdependent and nonlinear. In order to select the most important and interpretable features, we preselect the features by group Lasso with logistic loss.

Lasso (Least Absolute Shrinkage and Selection Operator) [11] provides a way to gain the sparsity of the parameters by imposing a 1-norm regularization. The objective function to be minimized is defined as

$$Q(\boldsymbol{\theta}) = loss(\boldsymbol{\theta}) + \lambda ||\boldsymbol{\theta}||_1, \tag{1}$$

where $\boldsymbol{\theta}$ is the parameter vector in the model, $loss(\boldsymbol{\theta})$ is the loss function, and $||\bullet||_1$ is the 1-norm. As λ is increased, the components of $\boldsymbol{\theta}$ are gradually shrunk to zero so as to achieve sparsity. The feature whose weight is shrunk later is considered to be more important, and so the shrinking order can be used to select features.

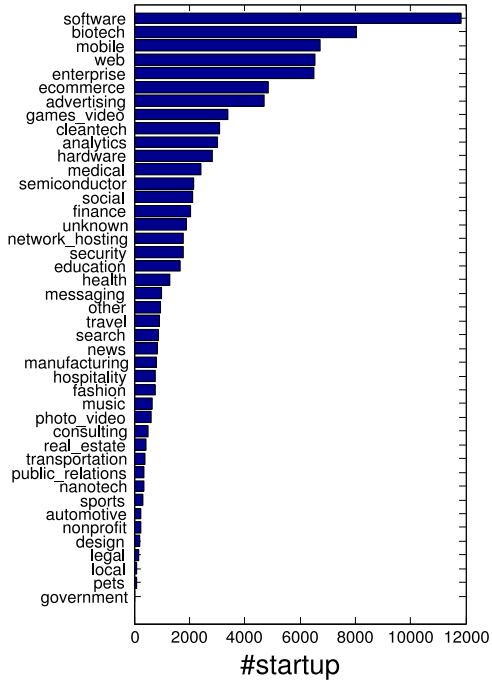


Fig. 7. Distribution of startup over field. Software, mobile web, and biotech are the most active fields for startup.

However, the categorical variable in the model is usually coded via dummy variables, and so the dummy variables corresponding to one categorical variable may be set to zero in different time in Lasso, which makes the sparsity of Lasso less powerful. Thus, Yuan and Lin [12] proposed group Lasso to shrink the dummy variables of a group together for least squared loss, and later, group Lasso was generalized to logistic loss [13]. The objective function is

$$Q(\boldsymbol{\theta}) = loss(\boldsymbol{\theta}) + \lambda \sum_{j=1}^{q} m_j ||\boldsymbol{\theta}_{G_j}||_2, \tag{2}$$

where all features are divided into j groups according to the coding of dummy variables, i.e., $G_1, G_2, \ldots, G_j$. The multiplier m_j serves for balancing cases where the groups are of different sizes, and $||\bullet||_2$ denotes the Euclidean norm.

We use a large number of categorical features, and group Lasso is employed to select features by group. Since the proposed SBFG (cf. Section 6) is a kind of generalized linear model, we choose the logistic loss for feature selection. We fit group Lasso for logistic regression by the group descent algorithm in the R package 'grpreg'[3] [14]. Group Lasso only needs the labeled training data, and the data in 1984-2010 of CRUNCH are fed to Group Lasso for determining the reverse shrinking order of features, which is shown in the third column of Table 2. For instance, the No. 12 feature has order 01, which means that the weight of the No. 12 feature is the last to shrink to zero, and so it is regarded as the most important feature in the model.

5 Observation of Prominent Features

As mentioned in Section 4, the features are categorized into three kinds, i.e., static features, dynamic domain features and dynamic topology features, and we will describe the top

3. http://cran.r-project.org/web/packages/grpreg/index.html

TABLE 2
All Features for Co-Investment

Feature	No.	Ord.*	Name	Short description
Static feature	01	43	latitudeMax	Larger value of latitude
	02	34	latitudeMin	Smaller value of latitude
	03	26	latitudeSingle	Single value of latitude
	04	33	latitudeDiff	Difference of latitude
	05	43	longitudeMax	Larger value of longitude
	06	31	longitudeMin	Smaller value of longitude
	07	16	longitudeSingle	Single value of longitude
	08	23	longitudeDiff	Difference of longitude
	09	20	absoluteDistance	Straight-line distance
	10	17	timeZoneDiff	Difference of time zone
	11	80	sameCity	Are two VCs in the same city?
	12	01	sameCountry	Are two VCs in the same country?
	13	14	ethnicitySim	Ethnicity similarity**
	14	39	languageSim	Language similarity**
	15	29	religionSim	Religion similarity**
	16	06	investorCombination	Combination of investor type
	17	36	sameCVCField	Are two company VCs of the same field?
Dynamic domain feature	18	08	fieldsMax	Larger value of #field
	19	07	fieldsMin	Smaller value of #field
	20	09	fieldsSingle	Single value of #field
	21	61	fieldsDiff	Difference of #field
	22	77	fieldsSum	Sum of #field
	23	10	fieldsJaccard	Jaccard similary of fields
	24	23	shortTrendMax	Larger value of short trend***
	25	13	shortTrendMin	Smaller value of short trend
	26	18	shortTrendSingle	Single value of short trend
	27	23	longTrendMax	Larger value of long trend
	28	39	longTrendMin	Smaller value of long trend
	29	12	longTrendSingle	Single value of long trend
	30	36	firstInvestYearMax	Larger value of first year of investment
	31	20	firstInvestYearMin	Smaller value of first year of investment
	32	81	firstInvestYearSingle	Single value of first year of investment
	33	61	firstInvestYearDiff	Difference of first year of investment
Dynamic topology feature	34	04	distanceBefore	Shortest distance of two VCs
	35	39	degreeMax	Larger value of degree
	36	52	degreeMin	Smaller value of degree
	37	20	degreeSingle	Single value of degree
	38	47	degreeDiff	Difference of degree
	39	69	degreeSum	Sum of degree
	40	36	shConstraintMax	Larger value of structural hole constraint****
	41	56	shConstraintMin	Smaller value of structural hole constraint
	42	34	shConstraintSingle	Single value of structural hole constraint
	43	73	shConstraintDiff	Difference of structural hole constratint
	44	67	shConstraintSum	Sum of structural hole constraint
	45	55	shConstraintMaxEgo	Larger value of structural hole constraint of ego net
	46	52	shConstraintMinEgo	Smaller value of structural hole constraint of ego net
	47	43	shConstraintSingleEgo	Single value of structural hole constraint of ego net
	48	71	shConstraintDiffEgo	Difference of structural hole constraint of ego net
	49	73	shConstraintSumEgo	Sum of structural hole constraint of ego net
	50	03	betweennessMax	Larger value of betweenness
	51	11	betweennessMin	Smaller value of betweenness
	52	05	betweennessSingle	Single value of betweenness
	53	76	betweennessDiff	Difference of betweenness
	54	42	betweennessSum	Sum of betweenness
	55	49	betweennessMaxEgo	Larger value of betweenness of ego net
	56	19	betweennessMinEgo	Smaller value of betweenness of ego net
	57	26	betweennessSingleEgo	Single value of betweenness of ego net
	58	65	betweennessDiffEgo	Difference of betweenness of ego net
	59	67	betweennessSumEgo	Sum of betweenness of ego net
	60	30	densityMaxEgo	Larger value of ego density
	61	56	densityMinEgo	Smaller value of ego density
	62	43	densitySingleEgo	Single value of ego density
	63	78	densityDiffEgo	Difference of ego density
	64	61	densitySumEgo	Sum of ego density

TABLE 2
(*Continued*)

Feature	No.	Ord.*	Name	Short description
	65	49	firstNeighborsMax	Larger value of #neighbor
	66	56	firstNeighborsMin	Smaller value of #neighbor
	67	26	firstNeighborsSingle	Single value of #neighbor
	68	47	firstNeighborsDiff	Difference of #neighbor
	69	72	firstNeighborsSum	Sum of #neighbor
	70	02	firstCommonNeighbors	#common neighbor
	71	65	secondNeighborsMax	Larger value of #secondary neighbor
	72	52	secondNeighborsMin	Smaller value of #secondary neighbor
	73	15	secondNeighborsSingle	Single value of #secondary neighbor
	74	69	secondNeighborsDiff	Difference of #secondary neighbor
	75	56	secondNeighborsSum	Sum of #secondary neighbor
	76	31	secondCommonNeighbors	#common secondary neighbor
	77	51	clusterCoefficientMax	Larger value of clustering coefficient
	78	73	clusterCoefficientMin	Smaller value of clustering coefficient
	79	56	clusterCoefficientSingle	Single value of clustering coefficient
	80	79	clusterCoefficientDiff	Difference of clustering coefficient
	81	61	clusterCoefficientSum	Sum of clustering coefficient

** Ord. denotes the reverse shrinking order of the feature in group Lasso, and the feature with a smaller Ord. is considered to be more important.*
*** Cf. [9] for calculation of similarity of ethnicity, language and religion.*
**** Follow-the-trend indicates that VCs tend to match their choices with the dominant choices of others, cf. [5] for more details.*
***** Cf. [10] for structural hole theory.*

10 features selected by group Lasso in this order, since the handling of static features is different from the dynamic ones.

5.1 Static Features

Nationality. Fig. 8 shows the distribution of VCs' nationality, where the orange bar denotes the positive instances (the existent co-investments in the dataset, cf. Section 7), and the blue bar denotes the negative instances. It is clearly shown that the two VCs tend to co-invest when they are from the same country.

Investor type. The investor types of VCs in CRUNCH are categorized into company venture capital (C), financial organization (F) and person investor (P). There are 2,875 company venture capitals, 8,038 financial organizations and 7,803 person investors in CRUNCH. Fig. 9 shows that financial organization tends to co-invest with financial organization, person investor tends to co-invest with person investor, and other combinations are not very popular.

5.2 Dynamic Domain Features

Invested fields reflect investment interest and investment diversity of VC. Since #field (# indicates the number, the same hereinafter) is a measure that changes over time, the values of the feature in different years are not directly comparable, and we use the rank of #field in the given year instead of the original value of #field, as show in Fig. 10. We employ the "equal-frequency binning" technique to discretize #field of the same year into a small number of distinct ranges. #field in the bin with the largest value is ranked 1, #field in the bin with the second largest value is ranked 2 and so on. After discretization within the same year, the comparability of #field in different years is improved. This technique is also applied to some other dynamic features, e.g. betweenness.

Since a potential co-investment involves two VCs, there are two values of #field for a potential co-investment. The larger one of the two values is shown in Fig. 10a, the smaller one is shown in Fig. 10b, and we find that VCs with large #field (small rank) tends to co-invest in both cases.

Jaccard similarity of invested fields. Besides the number of invested fields, we calculate the Jaccard similarity of the invested fields (jacc for short) for a VC pair, i.e.,

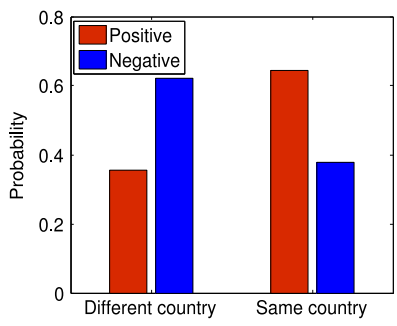


Fig. 8. Nationality. *Y*-axis: Probability, conditioned on nationality. The orange bar denotes the positive instances in the dataset, while the blue bar denotes the negative instances. Two VCs tend to co-invest when they are from the same country.

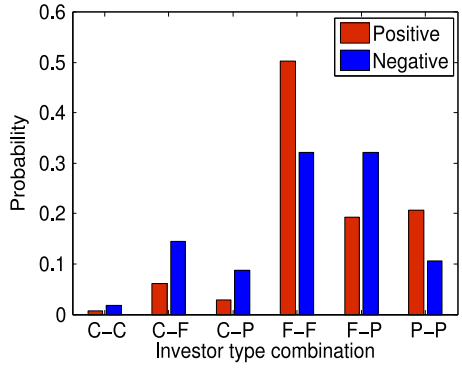
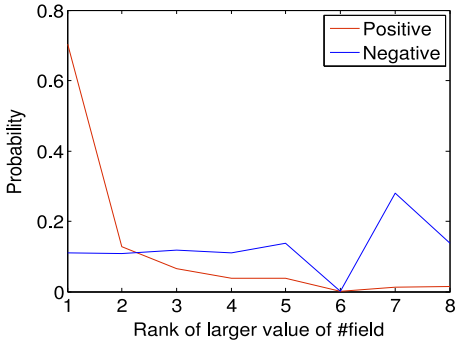
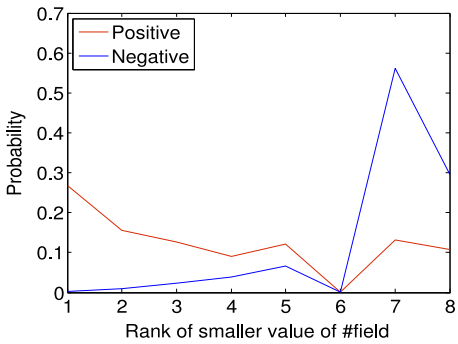


Fig. 9. Investor type combination. *Y*-axis: Probability, conditioned on investor type combination. "C" indicates company venture capital, "F" indicates financial organization, and "P" indicates person investor.



(a) Larger value



(b) Smaller value

Fig. 10. Invested fields. *Y*-axis: Probability, conditioned on rank of larger/smaller value of #field. Smaller rank indicates larger #field.

$\frac{|IF_t(vc_1)\bigcap IF_t(vc_2)|}{|IF_t(vc_1)\bigcup IF_t(vc_2)|}$, where $IF_t(vc_1)$ denotes the set of invested fields of vc_1 before time t. As shown in Fig. 11, when jacc is smaller than 0.1, the VC pair does not tend to co-invest, probably due to a lack of common interests. When jacc is larger than 0.8, the VC pair does not tend to co-invest either, probably because they cannot complement each other very well. Thus, the VC pair with appropriate jacc tends to co-invest.

5.3 Dynamic Topology Features

Besides the features from domain knowledge, there are also features related to social network that are selected by group Lasso, which are explained as follows.

Common neighbors reflect the link homophily between two VCs. Since this feature is related to the evolving VC network, we use the *common neighbor ratio* instead, which is defined as the ratio of the number of common neighbors to the sum of the number of neighbors of two VCs. Fig. 12 shows the histogram of positive instances and negative instances, where the bar heights are normalized so the area for each bar represents the probability for the corresponding interval. Comparing the histogram for positive instances (Fig. 12a) with the histogram for negative instances (Fig. 12b), we find that VCs with larger common neighbor ratio are more likely to co-invest with other VCs.

Betweenness is one of centrality measures of nodes in social network [15]. Although many centralities have been taken into consideration, such as degree, closeness and structural hole (cf. Table 2), betweenness is identified as the prominent feature in the co-investment prediction by group Lasso. Betweenness is a measure of the evolving VC network and so we use the rank of betweenness in the given year instead of the original value, as show in Fig. 13. Since a potential co-investment involves two VCs, there are two values of betweenness for a co-investment. The larger one of the two values is shown in Fig. 13a, the smaller one is shown in Fig. 13b, and we find that VCs with large betweenness (small rank) tend to co-invest in both cases.

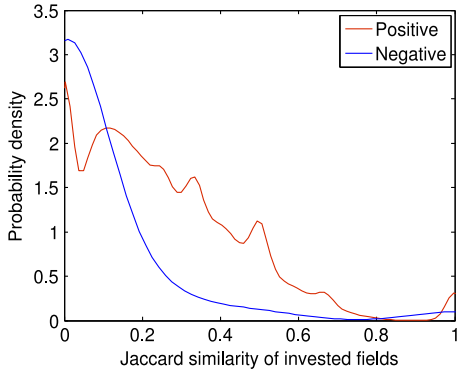
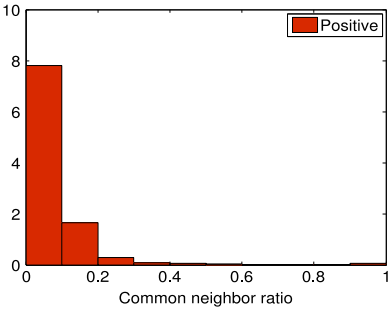
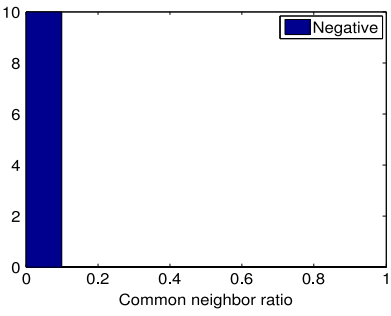


Fig. 11. Jaccard similarity of invested fields. The VC pair with appropriate (not too large and not too small) jacc tends to co-invest.

Shortest distance is considered to be one of the most important features in link prediction [8]. As shown in Fig. 14, when two VCs that have invested before (i.e., the shortest distance is 1) or have common neighbor (i.e., the shortest distance is 2), they are highly likely to co-invest. When the shortest distance is equal to or larger than 3, the likelihood of co-investment decreases rapidly. When there is no path between two VCs in the network (the shortest distance is **inf** in Fig. 14), i.e., one or two candidate VC(s) of the potential co-investment are not connected to the

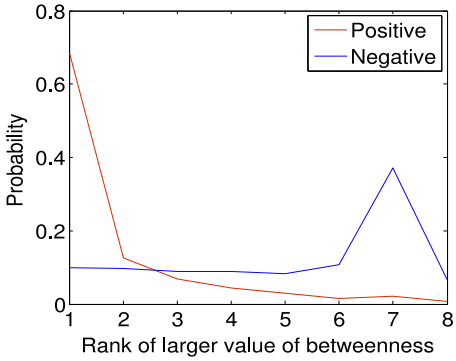


(a) Positive histogram

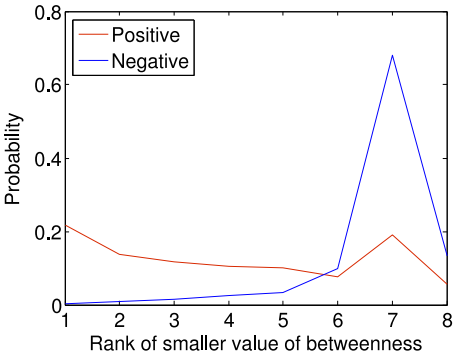


(b) Negative histogram

Fig. 12. Histogram of common neighbor ratio. The bar height is normalized so the area for each bar represents the probability for the corresponding interval.



(a) Larger value



(b) Smaller value

Fig. 13. Betweenness.Y-axis: Probability, conditioned on rank of larger/smaller value of betweenness. Smaller rank indicates larger betweenness.

biggest component of the VC network, they are not likely to co-invest.

6 Model Framework

Basically, the binary classification problem (co-invest or not) can be solved by any classifier, such as logistic regression and SVM. However, these models suffer from the same limitation that they cannot model the correlation between/among co-investments, so we try to develop an integrated factor graph model to capture both feature and correlation.

6.1 Structural Balance Theory

We explore an important pattern in the VC network based on the structural balance theory [15], which will be the theoretical foundation of our proposed model. Fig. 15 shows the

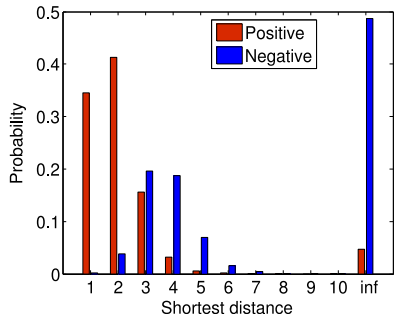


Fig. 14. Shortest distance on the VC network. VCs tend to co-invest when they have a short distance on the co-investment network.

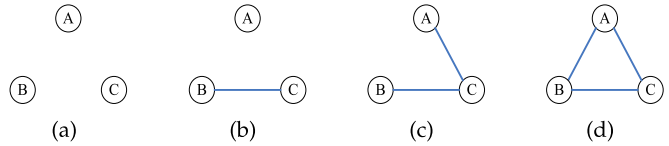


Fig. 15. Illustration of structural balance theory. The line between VCs indicates co-investment, and the two VCs connected by the line are called co-investors. The structural balance theory implies that either all three pairs of these VCs are co-investors (d) or only one pair of them are co-investor (b).

triad relationships, where the line between two VCs indicates the co-investment, and the two VCs connected by the line are called co-investors. For every group of three users (called triad), the structural balance theory implies that either all three pairs of these VCs are co-investors or only one pair of them are co-investor. As shown in Fig. 16, the number of balanced triads (those with three co-investments or one co-investments) is by far larger than that of unbalanced triads (those with two co-investments or zero co-investment) in CRUNCH. Moreover, the connected triads (those with two co-investments or three co-investments) are of particular interest to us, since the fact that the number of closed triads (those with three co-investments) is much larger than that of open triads (those with two co-investments) reflects the prevalence of triadic closure in the VC network, i.e., the co-investor of my co-investor is likely to be my co-investor.

6.2 The Proposed Model

The original VC network is built intuitively with VC as a node, but the goal of our research is to predict the co-investment between VCs. In addition, it is hard to model the correlation between/among co-investments (e.g. the triad correlation mentioned above) if with VC as a node. Thus, we prefer to model the co-investment as a node directly in the graphical model, and first the original VC network with VC as a node is converted to a graph model with co-investment as a node.

Our proposed model, i.e., structural balance based factor graph model, is inspired by the structural balance theory and observation in CRUNCH. The model is shown in Fig. 17. The left figure shows the original VC network, where the edges with label $1/0$ indicate whether two VCs co-invested or not in time span $\{1, t\}$, and the edges with label ? are those that we try to predict in time $t+1$. The solid/dashed line indicates whether the edge exists or not

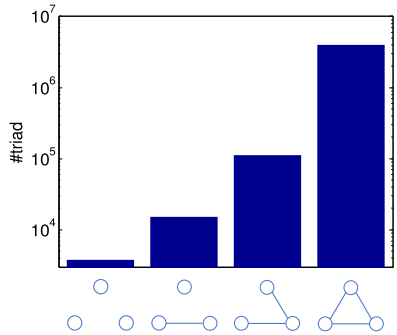


Fig. 16. Structural balance in the VC network. The number of balanced triads (those with three co-investments) is by far larger than that of unbalanced triads (those with two co-investments).

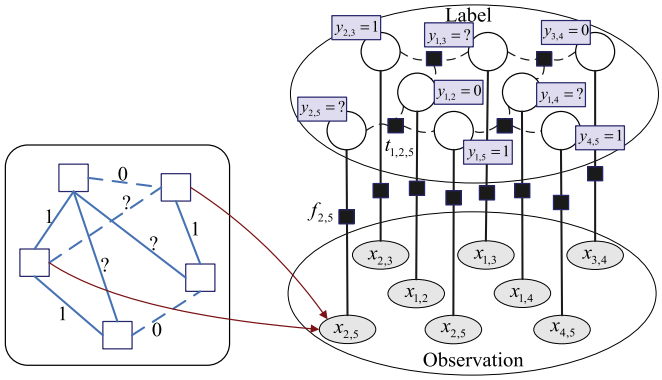


Fig. 17. Graphical representation of SBFG model. The left figure shows the original VC network with VC as a node, where the edges with label 1/0 represent whether two VCs co-invest or not in time span $\{1, t\}$, and the edges with label ? are those that we try to predict in time $t+1$. The right figure is the SBFG model with co-investment as a node. $y_{i,j}$ is the latent variable that indicates whether two VCs v_i, v_j co-invest, and $x_{i,j}$ is the observation of two VCs v_i, v_j. $f_{i,j}$ is the feature factor for a co-investment, and $t_{i,j,k}$ is the triad factor for three possible co-investments.

in the ground truth. The right figure is the SBFG model derived from the original VC network. $y_{i,j}$ is the latent variable that indicates whether two VCs v_1, v_2 co-invest or not, and $x_{i,j}$ denotes the observation of two VCs v_i, v_j. The SBFG model expresses the joint distribution over all variables as a product of factors over subsets of those variables, and the edge between a factor and a variable in the SBFG model indicates that the variable is an argument of the factor function. We define two kinds of factors in the SBFG model. $f(x_{i,j}, y_{i,j})$ ($f_{i,j}$ for short) represents the feature factor defined for the co-investment. $t(y_{i,j}, y_{i,k}, y_{j,k})$ ($t_{i,j,k}$ for short) represents the triad factor, which is used to capture the structural balance among three possible co-investments $y_{i,j}, y_{i,k}, y_{j,k}$ sharing common VCs.

Y denotes the vector that contains all latent variables. Since we already know the co-investments in time span $\{1, t\}$, the latent variable vector Y in the SBFG can be divided into labeled subset Y^L and unlabeled subset Y^U (to be predicted). We formalize the network with Markov random fields. According to Hammersley-Clifford theorem [16], the probability of latent variable vector Y given observations X can be factorized as

$$p(Y|X) = \frac{1}{Z} \prod_{possible\ i,j} f_{i,j} \prod_{possible\ i,j,k} t_{i,j,k}, \tag{3}$$

where "$possible\ i, j$" means all possible values that i, j can take in the dataset, and "$possible\ i, j, k$" has similar meaning. Factors are defined as

$$f_{i,j} = \exp\{\boldsymbol{\alpha}_{i,j}^T \mathbf{g}(x_{i,j}, y_{i,j})\} \tag{4}$$

$$t_{i,j,k} = \exp\{\beta_{i,j,k} h(y_{i,j}, y_{i,k}, y_{j,k})\}, \tag{5}$$

where $\mathbf{g}(x_{i,j}, y_{i,j})$ is the function vector for the feature factor $f_{i,j}$, $h(y_{i,j}, y_{i,k}, y_{j,k})$ is the function for the triad factor $t_{i,j,k}$, and $\boldsymbol{\alpha}_{i,j}$, $\beta_{i,j,k}$ are corresponding weights. The component of $\mathbf{g}(x_{i,j}, y_{i,j})$ is defined as

$$g_m(x_{i,j}, y_{i,j}) = \mathbf{1}_{\{y_{i,j}=\tilde{y}_{i,j}\}} \cdot g_m(x_{i,j}), \tag{6}$$

where $g_m(x_{i,j})$ is a certain feature for the observation $x_{i,j}$, as defined in Table 2. Each feature is nonzero only for a single label $\tilde{y}_{i,j}$. This particular form of function leads to a larger feature set, which is a common practice in feature engineering of Markov random fields, and can lead to better prediction accuracy since the final decision boundary can be more flexible [17]. The definition of $h(y_{i,j}, y_{i,k}, y_{j,k})$ is as follows:

$$h(y_{i,j}, y_{i,k}, y_{j,k}) = \mathbf{1}_{\{\#positive\ co-investment=a | a=0,1,2,3\}}. \tag{7}$$

That is to say, we use the number of positive co-investments in the triangle as a feature,[4] and there are a total of four features for the triad factor (if using indicator function form, as in Eq. (7)). When the model is fitted to the training data, the number of balanced triads is much larger than that of unbalanced ones, and so the weight of the feature for balanced triad should be larger than the weight of the feature for unbalanced triad after training. Finally, the model will encourage the balanced triads when predicting the test data.

Furthermore, we pack all weights $\boldsymbol{\alpha}_{i,j}$, $\beta_{i,j,k}$ into a long weighting vector $\boldsymbol{\theta}$, and pack all features $\mathbf{g}(x_{i,j}, y_{i,j})$, $h(y_{i,j}, y_{i,k}, y_{j,k})$ into a long feature vector $\mathbf{s}$, regardless of the type of factors. Thus, the conditional probability, i.e., Eq. (3), is simplified to be

$$p(Y|X) = \frac{1}{Z} \exp\{\boldsymbol{\theta}^T \mathbf{s}\}. \tag{8}$$

Then, we try to get proper weighting vector $\boldsymbol{\theta}$ in the learning phase.

6.3 Learning

The latent variables in time span $\{1, t\}$, i.e., Y^L, are labeled, and our optimization goal is to minimize the loss function, which is defined as the negative log-likelihood

$$\begin{aligned} -loss(\boldsymbol{\theta}) &= O(\boldsymbol{\theta}) \\ &= \log p(Y^L|X) = \log \sum_{Y^U} p(Y^L, Y^U|X) \\ &= \log \sum_{Y^U} p(Y|X) = \log \sum_{Y^U} \frac{1}{Z} \exp\{\boldsymbol{\theta}^T \mathbf{s}\} \\ &= \log \sum_{Y^U} \exp\{\boldsymbol{\theta}^T \mathbf{s}\} - \log Z \\ &= \log \sum_{Y^U} \exp\{\boldsymbol{\theta}^T \mathbf{s}\} - \log \sum_{Y} \exp\{\boldsymbol{\theta}^T \mathbf{s}\}. \end{aligned} \tag{9}$$

To minimize the loss function, we consider a gradient decent method, and the gradient is calculated as follows:

$$\frac{\partial O(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = E_{p(Y^U|Y^L,X)}[\boldsymbol{s}] - E_{p(Y^U,Y^L|X)}[\boldsymbol{s}], \tag{10}$$

where $E_{p(Y^U|Y^L,X)}[\boldsymbol{s}]$ and $E_{p(Y^U,Y^L|X)}[\boldsymbol{s}]$ are expectations of $\boldsymbol{s}$ on different distributions. The derivation of the two terms in the right part of Eq. (10) are similar, and we only present the former for abbreviation,

4. There is a slight abuse of the word "feature" here. The "feature" here represents the structure of a triad consisting of three co-investments (called "triad feature" temporarily), instead of the feature for only one co-investment as listed in Table 2 (called "node feature").

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left[\log\sum_{Y^U}\exp\{\boldsymbol{\theta}^T\mathbf{s}\}\right]=\frac{1}{\sum_{Y^U}\exp\{\boldsymbol{\theta}^T\mathbf{s}\}}\sum_{Y^U}\exp\{\boldsymbol{\theta}^T\mathbf{s}\}\cdot\mathbf{s}$$
$$=\sum_{Y^U}\frac{\exp\{\boldsymbol{\theta}^T\mathbf{s}\}}{\sum_{Y^U}\exp\{\boldsymbol{\theta}^T\mathbf{s}\}}\cdot\mathbf{s}=\sum_{Y^U}\frac{Z\cdot p(Y|X)}{\sum_{Y^U}Z\cdot p(Y|X)}\cdot\mathbf{s}$$
$$=\sum_{Y^U}\frac{p(Y^U,Y^L|X)}{p(Y^L|X)}\cdot\mathbf{s}=\sum_{Y^U}p(Y^U|Y^L,X)\cdot\mathbf{s}$$
$$=E_{p(Y^U|Y^L,X)}[\mathbf{s}] \tag{11}$$

The calculation of expectations in Eq. (10) is converted to the calculation of the marginal probability $p(Y^U|Y^L,X)$ and $p(Y^U,Y^L|X)$, and is further converted to message passing along edges in the graph, which can be done by the standard belief propagation [18]. When applied to tree-structured graph, the belief propagation gives the exact result. However, the graphical structure of our proposed SBFG can be arbitrary and contains cycles, and it's not feasible to use exact inference. We can still employ belief propagation to approximate the marginal probability, and the algorithm is called loopy belief propagation in this case [19]. Although the precise conditions of convergence of loopy belief propagation are not well understood [20], [21], it works well in our model. Note that we should perform LBP twice in each step, one for estimating marginal probability $p(Y^U,Y^L|X)$, and the other for $p(Y^U|Y^L,X)$. At the end of each step, we update the weighting vector $\boldsymbol{\theta}$ with the gradient and a constant learning rate η. η is set to 0.001, which is determined by preliminary experiments on a subset of the training data.

Algorithm 1. SBFG Learning Algorithm

Input: labeled variables Y^L, observations X, learning rate η
Output: weighting vector $\boldsymbol{\theta}$

```
Initialize θ;
while not converged do
    Calculate E_{p(Y^U|Y^L,X)}[s] using LBP;
    Calculate E_{p(Y^U,Y^L|X)}[s] using LBP;
    Calculate the gradient ∂O(θ)/∂θ according to Eq. (10);
    Update θ with θ^new = θ^old − η · ∂O(θ)/∂θ;
Return θ;
```

The learning algorithm is shown in Algorithm 6.3, and the time complexity of the algorithm is mainly determined by the computation of marginal probability using LBP. Generally, time complexity of LBP is $O(nES^C)$, where n represents the number of features, E the number of edges, S the number of labels, and C the size of the maximal clique.[5] In our case, $E=3T$, $S=2$, $C=3$, where T is the number of triads, and so the time complexity is $O(nT)$, which is linear function of the number of features n and the number of triads T.

6.4 Prediction

Once we get the learned weight vector $\boldsymbol{\theta}$, we can predict the unlabeled Y^U by first computing the marginal probability of $p(Y^U|Y^L,X)$ and then select the value with the largest marginal probability as the label. Again, the marginal probability of $p(Y^U|Y^L,X)$ is calculated by running LBP, and the marginal probability is then taken as the prediction confidence.

5. Although there are other factors that affect the time complexity of LBP (e.g. the number of iterations of gradient decent), their order of magnitude usually does not change much, so they are not included in the formula to facilitate the analysis.

TABLE 3
Prediction Performance of Co-Investment with the Top 10 Features

Data	Alg.	Pre.	Rec.	F1	Acc.
2011	SVC	**0.8615**	0.7082	0.7773	0.8078
	LR	0.8601	0.7071	0.7761	0.8068
	SBFG	0.8236	**0.9939**	**0.9008**	**0.8963**
2012	SVC	**0.8770**	0.7059	0.7822	0.8129
	LR	0.8721	0.7095	0.7825	0.8122
	SBFG	0.8431	**0.9939**	**0.9123**	**0.9090**
2013	SVC	**0.8693**	0.7124	0.7831	0.8104
	LR	0.8664	0.7133	0.7825	0.8095
	SBFG	0.8395	**0.9920**	**0.9094**	**0.9050**
2014 (first 3 months)	SVC	0.9143	0.7210	0.8062	0.8287
	LR	0.9164	0.7240	0.8089	0.8309
	SBFG	**0.9308**	**0.9924**	**0.9606**	**0.9598**
Average	SVC	**0.8805**	0.7119	0.7872	0.8150
	LR	0.8788	0.7135	0.7875	0.8149
	SBFG	0.8593	**0.9931**	**0.9208**	**0.9175**

Measures: Pre. denotes precision, Rec. denotes recall, F1 denotes F1 value, and Acc. denotes accuracy. Compared methods: SVC denotes support vector classifier, LR denotes logistic regression, and SBFG denotes the proposed method in this paper.

7 Experiments and Analysis

7.1 Experiment Setup

CRUNCH contains 18,716 VCs and 152,227 co-investment events from 1984 to 2014. The 152,227 co-investments are positive instances in our experiments. There are no direct negative instances in the dataset, and then we consider all possible combinations of accumulated VCs until a given time point. However, the number of combinations is hundreds of times larger than the number of positive instances, which constitutes imbalanced data. There a large amount of research on imbalanced data, and the methods include making the learning process active or cost-sensitive, and treating the classifier score with different thresholds [22], [23]. We employ random undersampling due to its effectiveness and ease of implementation. Specifically, we randomly sample the same number of negative instances as positive instances.

Our goal is to predict co-investments in time $t+1$ (test dataset), given data in time span $\{1,t\}$ (training dataset), and we construct four cases for CRUNCH. The first case is to predict co-investments in 2011 given 1984-2010, the second is 2012 given 1984-2011, the third is 2013 given 1984-2012, and the fourth is the first three months in 2014 given 1984-2013.

7.2 Prediction Performance

We compare our proposed model with state-of-the-art supervised machine learning algorithms, and the results are shown in Table 3.

Measures. For evaluating the prediction of co-investment, four popular measures are used to evaluate the performance, i.e., precision, recall, F1 measure and accuracy. Let *TP* denote #true positive, *FP* #false positive, *FN* #false negative and

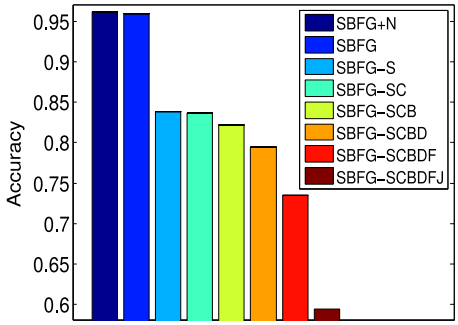


Fig. 18. Feature contribution analysis for the case of 2014. SBFG stands for the proposed method with the top 10 features. The plus mark denotes additional features besides the top 10 features, and the minus mark denotes features that are excluded from the top 10 features.

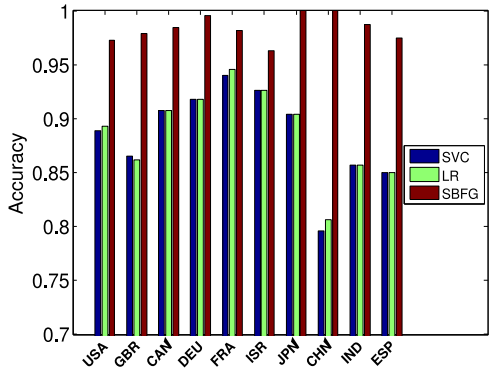


Fig. 19. Prediction performance of different countries. The performance gap between baseline method and SBFG of Asian countries is larger than other countries.

TN #true negative. $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$, $F1 = \frac{2*Precision*Recall}{Precision+Recall}$, and $Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$.

Baselines. The co-investment prediction is formulated as a binary classification problem in this paper. There are a large number of classifiers, and SVM and logistic regression are regarded as state-of-the-art general-purpose classifiers. The baselines are support vector classifier with L2 regularization (SVC) and logistic regression with L2 regularization (LR). These two algorithms are implemented in the LIBLINEAR software package [24]. All baselines use the top 10 features selected by group Lasso, but not the structural balance factor, since the point-wise classifiers (SVC and LR) cannot model the correlation among co-investments efficiently. SBFG employs both the top 10 features and the structural balance factor.

As shown in Table 3, SBFG significantly exceeds all state-of-the-art algorithms in all measures except precision. The prediction accuracy and F1 value of SBGF are above 0.9, which are satisfactory for co-investment prediction.

7.3 Feature Contribution Analysis

We examine the contribution of different features by removing them one by one in the model for the case of 2014. As shown in Fig. 18, SBFG stands for the proposed method with the top 10 features. The plus mark denotes additional features besides the top 10 features, and the minus mark denotes features that are excluded from the top 10 features. N denotes the remaining 71 features other than the top 10 features, S the structural balance factor, C common neighbors, B betweenness, D shortest distance, F the number of invested fields, and J the Jaccard similarity of invested fields.

When the 71 remaining features are excluded from the model, the accuracy drops by only 0.18 percent (from 96.16 to 95.98 percent), which shows that the top 10 features selected by group Lasso can explain the formation of the VC network quite well. Note that, if there is no feature selection mechanism like group Lasso, it is hard to say that betweenness centrality is more predictive than structural hole constraint in co-investment prediction. When the structural balance factor is removed from the model, the accuracy drops by 12.5 percent (from 93.85 to 83.85 percent), which demonstrates the prediction power of structural balance theory. When the features are excluded from the model one by one, the performance drops gradually. Finally, when all dynamic features SCBDFJ are excluded from the model, there are only two static features, i.e., nationality and investor type, and the accuracy of the model is only 59.38 percent, which is slightly better than a random guess since the task is predicting the evolving link formation.

7.4 Country Analysis of Prediction

We analyze the prediction performance for the top 10 countries with the most VCs. We calculate accuracy for co-investments that involve the given country respectively, as shown in Fig. 19. It is shown that the proposed method SBFG exceeds the baselines by a large margin for all 10 countries. The average accuracy of baselines for Asian countries (Japan, China and India) is relatively low compared with other countries, while the average accuracy of SBFG for Asian countries is not low. Furthermore, China is the country with the lowest baselines accuracy and the highest SBFG accuracy. It probably suggests that VCs of Asian countries, especially of China, are more likely to have social relations due to their special economic culture, and they rely on the robustness of networks to avoid risks.

7.5 Investor Type Analysis of Prediction

We analyze the prediction performance for different investor types by calculating accuracy for co-investments that involve the given investor type, as shown in Fig. 20. *FinanOrg* denotes financial organization, *Company* denotes company investor, and *Person* denotes person investor. For the baseline algorithms (SVC and LR), the accuracy of person investor is by far lower than (-20 percent) financial organization and company investor. However, our proposed SBFG model can largely compensate for the performance

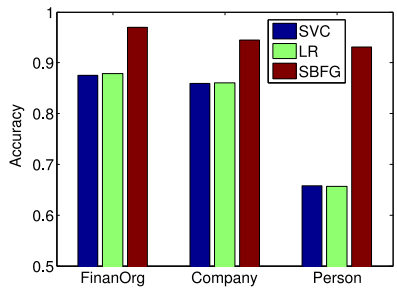


Fig. 20. Investor type analysis. *FinanOrg* denotes financial organization, *Company* denotes company investor, and *Person* denotes person investor.

(a) SVC/LR (b) SBFG

Fig. 21. Case study. The goal is to predict co-investments in 2014 given data in 1984-2013. SVC and LR misses two co-investments, while SBFG successfully predicts them by incorporating structural balance theory.

gap for person investor (only 4 percent lower than financial organization), which demonstrates the power of structural balance in co-investment prediction.

7.6 Case Study

Now we present a case study to demonstrate the effectiveness of the proposed model. In Fig. 21, each node represents a VC. The node in the upper left corner is Draper Fisher Jurvetson (DFJ for short), upper right Nexus Venture Partners (Nexus), lower left Gray Ghost Ventures (Gray) and lower right Garage Technology Ventures (Garage). The line between nodes denotes the co-investment, and the mark on the line indicates that the algorithm makes a mistake. Our goal is to predict co-investments in 2014 given 1984-2013. SVC and LR correctly predict the co-investments of DFJ-Gray and Nexus-Garage, but they miss the other two ones. Besides DFJ-Gray and Nexus-Garage, our proposed SBFG successfully predicts Gray-Nexus and Gray-Garage. After adding Gray-Nexus and Gray-Garage, both the upper left triangle and the lower left triangle become balanced.

8 Study on another Dataset

Another investment dataset (CHN[6]), which focuses on investments to Chinese startups, is also explored to further verify the model and features mentioned above.

8.1 Data Description

It takes two years to collect and clean the data in CHN manually, and then verify it by a questionnaire. CHN contains investments for Chinese startup from 1995 to 2011, and there are a total of 1,541 VCs, 5,494 Chinese startups, 10,275 investments and 5,856 co-investments. 50.8 percent of investments in CHN are related to two or more investors, which is much lower than that of CRUNCH (80.9 percent). CHN is five times (in terms of the number of investments) larger than the subset of CRUNCH related to Chinese startups (denoted by CRUNCH-China), although the time range of CHN is only half of CRUNCH-China. The information and statistics of CHN and CRUNCH-China are summarized in Table 4.

The different names of information between CHN and CRUNCH are italicized in Table 4. For example, CHN provides property right and year of establishment for VC firms. Besides the information of different names, the information with the same name could also be different between CHN and CRUNCH-China. For example, VCs of CHN are categorized into seven investor types, i.e., angle, venture capital, private equity, company venture capital, strategic investor, bank/trust, and other, which is different from the categorization of the three investor types of CRUNCH. In addition, there are 20 coarse-grained fields and 205 fine-grained fields in CHN, which is also different from the categorization of 44 fields of CRUNCH. Although the genre, size and information of CHN is different from CRUNCH, the structural balance phenomenon also holds in CHN (the pattern is quite similar to Fig. 16 and so omitted).

6. We will publish this dataset on the publicly available website after being anonymized.

TABLE 4
Information and Statistics of CHN and CRUNCH-China

Item	CHN (1995-2011)	CRUNCH-China (1984-2014)
Investment information	VC, Startup, Funded year	VC, Startup, Funded year, *Round*, *Rasied amount*
VC information	Investor type, Location, *Property right*, *Year of establishment*	Investor type, Location, *Field*
Startup information	Field, Location	Field, Location
#Investment	10,275	1,986
#VC	1,541	453
#Startup	5,494	781

8.2 Performance and Discussion

We employ the top 10 features in Table 2, plus the extra information provided by CHN, i.e., property right, year of establishment and features related to fine-grained fields, to train the model and make predictions. We construct four datasets that are with settings similar to CRUNCH (cf. Section 7). The experiment results are shown in Table 5.

Due to the different genre, size and information, the results of CHN and CRUNCH are not directly comparable. However, the accuracy of our proposed SBFG model is over 90 percent on CHN, and it outperforms SVC or LR significantly (+12 percent in accuracy), which further verifies the effectiveness of the proposed model and features.

TABLE 5
Prediction Performance of Co-Investment in CHN

Data	Alg.	Pre.	Rec.	F1	Acc.
2008	SVC	0.7729	0.7015	0.7354	0.7565
	LR	0.7925	0.7363	0.7634	0.7797
	SBFG	**0.8627**	**0.9505**	**0.9045**	**0.9031**
2009	SVC	0.7791	0.7556	0.7672	0.7784
	LR	0.7869	0.7218	0.7529	0.7711
	SBFG	**0.9136**	**0.8741**	**0.8934**	**0.8992**
2010	SVC	0.8444	0.6609	0.7415	0.7764
	LR	0.8472	0.7093	0.7721	0.7968
	SBFG	**0.9212**	**0.9006**	**0.9108**	**0.9144**
2011	SVC	0.7944	0.7385	0.7654	0.7814
	LR	0.7920	0.7385	0.7643	0.7801
	SBFG	**0.8887**	**0.9203**	**0.9042**	**0.9059**
Average	SVC	0.7977	0.7141	0.7524	0.7732
	LR	0.8047	0.7265	0.7632	0.7819
	SBFG	**0.8966**	**0.9114**	**0.9032**	**0.9057**

9 Related Works

9.1 Co-investment

In sociology and economics, the study of co-investment dates back to Wilson's theory on syndication [25], and Lerner [2] studied the principle of who will be a good co-investor and when to reconstruct a co-investment. More recently, some scholars studied co-investment/syndication from the perspective of link formation, such as [3], [4], [5], [26]. Based on 45 years of VC data from the US, [4] found several features that might have influence on the new link. However, [4] only used the node features and they did not make predictions. Powell et al. [5] studied four kinds of effects on interorganizational collaboration. The existing researches only explored a few features for co-investment without detailed analysis of contribution of features.

9.2 Link Prediction

Our work is related to link prediction, and the existing works on link prediction can be broadly grouped into two categories based on the learning algorithms: unsupervised link prediction and supervised link prediction. The classic works of unsupervised prediction were surveyed in [7] and recently [27] designed a flow based method. There are many works on supervised link prediction, such as [6], [28], [29], [30], [31], [32]. [29] studied the extent to which the formation of a reciprocal relationship can be predicted in a dynamic network. [30] developed a framework for classifying the type of social relationships by learning across heterogeneous networks. The co-investment network is intrinsically dynamic and multi-dimensional, and there is still nothing reported about the prediction of co-investment as far as we are informed. In this work, we focus on studying the underlying patterns that influence the formation of co-investment and propose a factor graph model to incorporate structural balance theory and the discovered patterns.

10 Conclusion and Future Work

In this paper, we study the prediction of co-investment of VCs. We present a series of observation analysis, design a large number of features, and then select prominent features for co-investment by group Lasso. Then we propose a factor graph model SBFG based on structural balance theory to formalize the observation into a unified model. For the model learning, we employ the loopy belief propagation to obtain an approximate solution. Experiment results show that the proposed method can accurately (around 90 percent in terms of accuracy) predict the co-investment in the near future with only 10 features selected by group Lasso, and obtains a significant improvement (+9 percent in terms of accuracy) over the baselines.

In the future, we will further explore VC investment in the following directions. First, we will design a SBFG model with an embedded feature selection mechanism, which can better explain the formation of the VC network and further improve the prediction performance. Second, although the proposed model exceeds consistently the baselines in most measures, the precision varies for different datasets, and we plan to study the effect of different datasets on the proposed model. Last, we will study other structural patterns that may affect the formation of VC network, such as circle with more than three nodes.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61303068, and the Research Fund of State Key Laboratory of High Performance Computing under Grant No. 201502-02. The authors would like to thank Jing Zhang, Huaiyu Wan, Zhanpeng Fang, and Ling Zhou at Tsinghua University for their help. Y. Zhou is the corresponding author.

References

[1] D. Trpido, "Mechanisms of venture capital co-investment networks: Evolution and performance implications," *Unpublished manuscript*, 2009, http://web.stanford.edu/group/esrg/siliconvalley/docs/coinvestments_DTrapido.pdf

[2] J. Lerner, "The syndication of venture capital investments," *Financial Manage.*, vol. 23, no. 3, pp. 16–27, Autumnd 1994.

[3] O. Sorenson and T. Stuart, "Syndication networks and the spatial distribution of venture capital investment," *The Am. J. Sociol.*, vol. 106, no. 6, pp. 1546–1588, May 2001.

[4] B. Kogut, P. Urso, and G. Walker, "Emergent properties of a new financial market: American venture capital syndication, 1960-2005," *Manage. Sci.*, vol. 53, no. 7, pp. 1181–1198, Jul. 2007.

[5] W. Powell, D. White, K. Koput, and J. Owen-Smith, "Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences," *Am. J. Sociol.*, vol. 110, no. 4, pp. 1132–1205, Jan. 2005.

[6] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2011, pp. 635–644.

[7] D. Liben-Nowell and J. Kleinberg, "The Link-prediction problem for social networks," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 7, pp. 1019–1031, May 2007.

[8] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *Proc. SDM Workshop Link Anal., Counterterrorism Security*, 2006, http://www.siam.org/meetings/sdm06/workproceed/Link%20Analysis/12.pdf

[9] A. Alesina, A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg, "Fractionalization," *J. Econ. Growth*, vol. 8, no. 2, pp. 155–194, Jun. 2003.

[10] R. Burt, *Structural Holes: The Social Structure of Competition*. Cambridge, MA, USA: Harvard Univ. Press, 1992.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc., Series B*, vol. 58, pp. 267–288, 1996.

[12] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc., Series B*, vol. 68, pp. 49–67, 2006.

[13] L. Meier, S. van de Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Statist. Soc., Series B*, vol. 70, pp. 53–71, 2008.

[14] P. Breheny and J. Huang, "Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors," *Statist. Comput.*, vol. 25, pp. 173–187, Mar. 2015.

[15] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[16] J. Hammersley and P. Clifford, "Markov field on finite graphs and lattices," *Unpublished manuscript*, 1971, http://www.statslab.cam.ac.uk/~grg/books/hammfest/hamm-cliff.pdf

[17] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Found. Trends Mach. Learn.*, vol. 4, no. 4, pp. 267–373, 2011.

[18] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA, USA: Morgan Kaufmann, 1988.

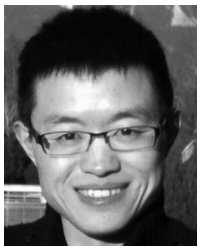
[19] B. Frey and D. MacKay, "A revolution: Belief propagation in graphs with cycles," in *Proc. Conf. Neural Inf. Process. Syst.*, 1997, pp. 479–485.

[20] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural Comput.*, vol. 12, no. 1, pp. 1–41, Jan. 2000.

[21] J. M. Mooij and H. J. Kappen, "Sufficient conditions for convergence of the Sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4422–4437, Dec. 2007.
[22] N. V. Chawla, N. Japkowicz, and A. Kołcz, "Editorial: Special issue on learning from imbalanced data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 1–6, 2011.
[23] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[24] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Aug. 2008.
[25] R. Wilson, "The theory of syndicates," *Econometrica*, vol. 36, no. 1, pp. 119–132, Jan. 1968.
[26] M. Piskorski, "Networks of power and status: Reciprocity in venture capital syndicates," *Unpublished manuscript*, 2004, https://scholar.google.com/citations?view_op=citation&hl=en&user=6txpXI8AAAAJ&citation_for_view=6txpXI8AAAAJ:d1gkVwhDpl0C
[27] R. Lichtenwalter, J. Lussier, and N. Chawla, "New perspectives and methods in link prediction," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 243–252.
[28] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proc. Int. Conf. World Wide Web*, 2010, pp. 641–650.
[29] J. Hopcroft, T. Lou, and J. Tang, "Who will follow you back? reciprocal relationship prediction," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 1137–1146.
[30] J. Tang, T. Lou, and J. Kleinberg, "Inferring social ties across heterogenous networks," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2012, pp. 743–752.
[31] Y. Yang, J. Tang, J. Keomany, Y. Zhao, J. Li, Y. Ding, T. Li, and L. Wang, "Ming competitive relationships by learning across heterogeneous networks," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1432–1441.
[32] S. Wu, J. Sun, and J. Tang, "Patent partner recommendation in enterprise social networks," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2013, pp. 43–52.

Zhiyuan Wang received the BSc, MSc, and PhD degrees from the National University of Defense Technology in 2003, 2005, and 2011, respectively. She is now an assistant professor at the State Key Laboratory of High Performance Computing, National University of Defense Technology and the School of Computer, National University of Defense Technology. Her research interests focus on data mining, parallel and distributed systems, and robotics. She is a member of the IEEE.

Yun Zhou is now an assistant professor in the State Key Laboratory of High Performance Computing, National University of Defense Technology and the School of Computer, National University of Defense Technology. His research interests focus on machine learning, natural language processing, social network analysis, and robotics.



Jie Tang is an associate professor at Tsinghua University. His research interests are social network analysis, data mining, and semantic web. He is a senior member of the IEEE.

Jar-Der Luo is a full professor at Tsinghua University. His research interests are social network analysis and economics.