# Predição de Relacionamentos em Redes Sociais, uma Revisão Sistemática

**William Takahiro Maruyama**[1], **Luciano Antonio Digiampietri**[1]

[1]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP)
Av. Arlindo Béttio, Ermelino Matarazzo – 03828-000 – São Paulo – SP – Brasil

***Abstract.*** *The social network analysis area is on the rise. An important task in this area is link prediction, in which the goal is to predict connections between users. For this task it is necessary the use of attributes, methods, algorithms and techniques that measure, somehow, the possibility of a relationship be created. However, there are many approaches and combinations of attributes for predicting links. This paper aims to conduct a comprehensive survey of the attributes/characteristics that can be used to predict links in various contexts of social networks, based on the Systematic Review methodology.*

***Resumo.*** *A área de análise de redes sociais está em ascensão. Uma importante tarefa desta área é a predição de relacionamentos, na qual o objetivo é prever conexões entre usuários. Para a realização desta tarefa são utilizados atributos, métodos, algoritmos e técnicas que medem, de alguma forma, a possibilidade de um relacionamento ser criado. No entanto, existem muitas abordagens e combinações de atributos para predizer relacionamentos. Este trabalho tem como objetivo realizar um levantamento abrangente dos atributos ou características que podem ser utilizados na predição de relacionamentos nos diversos contextos das redes sociais, a partir da metodologia de Revisão Sistemática.*

## 1. Introdução

Cada vez mais presente no dia-a-dia das pessoas, as redes sociais online (tais como Facebook, LinkedIn, Google+, Twitter, Flickr, etc) são compostas por inúmeros indivíduos, os quais podem estabelecer entre si diversos tipos de interação ou relação (genericamente conhecidas como *links*). Em geral, as redes online crescem continuamente e expressivamente, além disso, muitas delas são enormes e esparsas. Adicionalmente, elas são naturalmente dinâmicas, pois a todo instante novas relações são estabelecidas ou desfeitas [Shin et al. 2012, Song et al. 2009, Song et al. 2012, Vasuki et al. 2011]. Neste contexto, este trabalho objetiva o levantamento das referências que estudam a dinâmica dos *links* em uma rede social, focando nos trabalhos que visam a prever novos relacionamentos que são mais prováveis de ser formados. Essa tarefa, da área Análise de Redes Sociais, é conhecida como Predição de Relacionamentos ou de Ligações *(Link Prediction)* [Liben-Nowell and Kleinberg 2003, Hasan et al. 2006, Murata and Moriyasu 2008, Hasan and Zaki 2011, Lu and Zhou 2011, Dhote et al. 2013].

A predição de novos relacionamentos dentro de uma rede social é uma tarefa que ganhou bastante destaque nos últimos anos e serve para, desde encontrar amigos que ainda não estavam ligados em numa rede social *online* [Vasuki et al. 2010, Tian et al. 2010, Perez et al. 2012, Fire et al. 2011,

Zhong et al. 2013, Quercia and Capra 2009], até para potencializar a realização de trabalhos em empresas ou na academia [Maruyama and Digiampietri 2016, Digiampietri et al. 2015, Digiampietri and Maruyama 2014, Digiampietri et al. 2013, Hsieh et al. 2013, Dong et al. 2012, de Sa and Prudencio 2011].

Para realizar a predição utiliza-se de métodos que de alguma maneira medem a proximidade ou similaridade entre as entidades (nós) da rede. Esses métodos fornecem medidas que podem ser utilizadas por si só para predizer, mas podem ser adotadas como atributo (ou característica) a serem utilizados por uma estratégia de predição supervisionada [de Sa and Prudencio 2011, da Silva Soares and Bastos Cavalcante Prudencio 2012]. A identificação dos melhores conjuntos de atributos relevantes, dentre as combinações possíveis, é de muita importância para a melhoria de precisão dos modelos preditivos.

## 2. Metodologia e Condução

Com base em uma pesquisa exploratória realizada preliminarmente, foram identificadas *Link*, *co-autorship*, *prediction*, *social networking* e *scientific collaboration networking* como as principais palavras-chaves relacionadas ao assunto. Posteriormente, através da metodologia de revisão sistemática foi criado o protocolo que define e formaliza os procedimentos seguidos no presente estudo [Biolchini et al. 2005].

Este artigo tem como objetivo responder a seguinte pergunta: quais atributos existentes estão sendo utilizados na predição de coautorias em Redes Sociais? Para responder a esta questão foram feitas pesquisas nas principais bibliotecas digitais científicas da área, as quais disponibilizam os trabalhos via *web*. As bibliotecas digitais utilizadas foram: IEEExplore Digital Library e ACM Digital Library.

Embora seja possível encontrar outros artigos sobre o assunto dispersados em diferentes revistas ou anais de eventos, optou-se pela consulta apenas a estas duas bibliotecas digitais por serem consideradas as que congregam um maior número de artigos na área.

Com base nas palavras-chaves selecionadas foram criadas e submetidas as expressões e opções de busca em cada uma das bibliotecas. Objetivou-se tanto encontrar artigos sobre predição de relacionamentos em qualquer tipo de rede social, assim como predição de relacionamentos em redes de colaboração científica. Para não restringir o resultado da busca, não foi considerado um período de publicação. A chave de busca submetida à biblioteca digital da ACM foi "*((Abstract:"Link" OR Abstract:"co-authorship") AND (Abstract:"Prediction") AND (Abstract:"social network" OR Abstract:"scientific collaboration network"))*", além disso foi ativada a busca avançada, com utilização apenas do campo *abstract*. A chave de busca submetida à biblioteca digital IEEExplore foi " *((Abstract:"Link" OR Abstract:"co-authorship") AND (Abstract:"Prediction") AND (Abstract:"social network" OR Abstract:"scientific collaboration network"))*", além disso foi ativada a busca avançada, com filtro "Metadata only" ativo.

Todos os artigos encontrados na busca foram avaliados e selecionados segundo os critérios de inclusão e de exclusão que se seguem. Para aceitação do artigo, ele deve se enquadrar em todos os critérios de inclusão e nenhum de exclusão.

**Critérios de inclusão:** (i) Serão incluídos trabalhos completos publicados e disponíveis integralmente nas bases de dados científicas utilizadas. (ii) Serão incluídos trabalhos que analisem Redes Sociais. **Critérios de exclusão:** (i) Serão excluídos trabalhos de estudos secundários. (ii)

Serão excluídos trabalhos que não discutam os atributos que foram usados ou como foram usados para a predição de *links*. (iii) Serão excluídos trabalhos publicados que não estejam disponíveis integralmente nas bases de dados científicas especificadas.

Os resultados obtidos das duas etapas do protocolo da revisão sistemática estão organizados a seguir. O próximo parágrafo apresenta o resultado da triagem dos artigos (condução). Enquanto que a Seção 3 contém os resultados da sumarização das informações dos artigos incluídos na condução, análise das informações extraídas e a descrição de alguns artigos.

Com a submissão das expressões em cada uma das *engines* de busca das bibliotecas, foram encontrados inicialmente: 39 artigos na ACM e 37 artigos na IEEE. Desses artigos, ocorreu apenas um caso de repetição. Portanto, 75 artigos foram analisados. Uma seleção inicial foi realizada a partir dos critérios de inclusão e exclusão, aplicados sobre a leitura dos resumos (*abstracts*) de cada artigo. Nesta etapa, 12 artigos encontrados na ACM foram rejeitados e 14 na IEEE.

## 3. Extração

Os 49 artigos incluídos nesta revisão, foram lidos na íntegra e suas principais informações foram extraídas. Além dos dados bibliográficos, a Tabela 1 sumariza as informações extraídas de cada artigo, levando-se em consideração o foco do presente trabalho.

Com o levantamento realizado, foi observado que as publicações sobre este assunto são recentes. Foram incluídos: 1 artigo de 2003, 1 artigo de 2007, 3 artigos de 2009, 6 artigos de 2010, 15 artigos de 2011, 17 artigos de 2012 e 6 artigos de 2013. Nota-se que de 2010 a 2011 ocorreu um aumento de aproximadamente 71% nas publicações sobre assunto. Além disso, é nos últimos três anos que se concentra cerca de 77% das publicações. Portanto, é possível observar que este assunto tem, recentemente, atraído a atenção da comunidade científica da área.

Ao analisar a distribuição geográfica das publicações e tomando como base os dados de localização do primeiro autor, os resultados mostram que as pesquisas nessa área se concentram nos EUA, com 21 publicações, seguido pela China com 10. O Brasil localiza-se em quarto lugar com 3 publicações. Sobre os conjuntos de dados utilizados nos 49 artigos incluídos neste trabalho, foram registradas 57 fontes de dados diferentes (das 107 encontradas). Para tal análise, deve-se levar em consideração que um artigo pode ter utilizado mais de uma base de dados. A quantidade de fontes demonstra que há grande variedade de domínios.

Dentre todos os 49 artigos, o primeiro artigo publicado encontrado sobre a predição de *links* foi o de [Liben-Nowell and Kleinberg 2003], no qual os autores propuseram a predição de arestas (*links*) futuras com base nas arestas atuais, utilizando diversas medidas de proximidade (atributos) de nós em uma rede. Para tal, os autores utilizaram conjuntos de dados do arXiv, nos quais realizaram a predição de *links* de coautoria. Como resultado, eles concluíram que o atributo Katz e as variantes utilizadas apresentaram bom desempenho na maioria dos conjuntos de dados, sendo os melhores resultados obtidos em três dos cinco conjuntos. Além disso, segundo os autores, os atributos simples como CN e AA apresentaram resultados satisfatórios.

Em [Gao et al. 2011], os autores propuseram um modelo unificado de múltiplas informações da rede para predizer *links* de coautoria em função do tempo. Essas informações são de três tipos: da estrutura global da rede, o conteúdo dos nós e as informações de proximidade nos grafos para capturar os padrões de evolução ao longo do tempo das ligações nas redes. Utilizando quatro conjuntos de dados do arXiv entre 1992 e 2002, os resultados apresentados demonstram, conforme os autores, que o método proposto é eficiente em vários conjuntos de dados, podendo, de acordo com os valores de *Area Under Curve* (AUC), superar os métodos tradicionais para predição de *links* temporais. Os autores comentam também a possibilidade do uso do modelo proposto em redes de larga escala.

**Tabela 1. Tabela de extração dos dados**

| Referência | Base de dados | Atributos utilizados | Domínio de aplicação |
|---|---|---|---|
| [Aiello et al. 2012] | Last.fm e aNobii | Informações do perfil do usuário | Predição de *links* de amizade |
| [Almansoori et al. 2011] | Matriz com 24 encaminhamentos médicos | *Ethnicity* (E), *Professional Activity Match* (PAM), *Sum of Patients* (SofP), *Sum of Neighbors* (SofN) e *Jaccard Similarity* (JS ou JC de *Jaccard Coefficient*) | Predição de *links* positivos entre médicos |
| [Chang and Yao 2011] | Enron Email | *Singular value decomposition* (SVD), *Affinity measure* (AF) | Predição de *links* de trocas de e-mail |
| [Chelmis and Prasanna 2012] | Serviço de microblog corporativo (semelhante ao Twitter) | *Random, Shortest Distance*, Common Neighbors (CN), *Shared Vocabulary*, *Uniform Weighting Scheme* (SS_Uniform) | Predição *links* de intenção de comunicação |
| [Corlette and Shipman 2010] | Live Journal | Adamic-Adar (AA) e *Local Clustering Coefficient* | Predição de *links* de amizade (com efeito da abertura da rede) |
| [Costa and Ortale 2012] | Small World network e Enron Email | Bayesian Hierarchical Community-and-Role Model (BH-CRM), Latent Dirichlet Allocation for Graphs (LDA-G.) | Predição de *links* de interações de e-mails e citações |
| [Cukierski et al. 2011] | Flickr | Katz, CN, Adamic-Adar (AA), Cosseno, *Preferential Attachment* (PA), *Bayesian Sets*, SVD *Features*, SimRank, EdgeRank, *Commute Time*, *Bounded WalkPageRank*, *Maximum Flow*, *Betweenness Centrality*, *Core Number*, *Shortest Paths Histogram*, *Power Law Exponent* | Predição de *links* para separar relacionamentos reais de falsos |
| [Dong et al. 2012] | Epinions, Slashdot, Wikivote e Twitter | CN, AA, JC, PA, *ranking factor graph*, *out-degree*, *in-degree* e *all-degree* | Predição de *links* de interações em rede homogêneas e heterogênea |
| [Dong et al. 2011a] | CDRs de duas operadoras anônimas em uma cidade | CN, AA, JC, PA, *Hub Promoted Index* (HPI), *Hub Depressed Index* (HDI), Salton *Index* (SA),*Unweighted Random Walk* (URW), *Weighted call times random walk* (TRW), *Weighted call duration random walk* (DRW), *High-Performance Link Prediction* (HPLP), *Resource allocation based on weighted random walk* (RAURW), *Resource allocation based on weighted call times random walk* (RATRW), *Resource allocation based on weighted call duration random walk* (RADRW) | Predição de *links* de chamadas |
| [Dong et al. 2011b] | Power grid Network (PG), Political blogs network (PB), High-energy theory collaborations Network (Hep-th), Alex Arenas's Jazz, Alex Arenas's Email Network, Neural network of Elegans Network e US Air Network | CN, SA, Leicht-Holme-Newman *Index* (LHN), Sorensen *Index* (SOR), JC, HPI, HDI, PA, AA | Predição de *links* de interação em diversos tipos de redes |
| [Fire et al. 2011] | Academia, Facebook, Flickr, TheMarker e YouTube | *Vertex degree features*, *Vertex subgraphs features*, CN, *Total-Friends*, JC, *Transitive Friends*, PA, Katz, *Friends-measure*, *Opposite direction friends*, *Edge subgraphs edges number*, *Edge subgraphs components number*, *Shortest Path* (SP) | Predição de *links* faltantes de relacionamento em redes direcionadas e não direcionadas |
| [Gao et al. 2011] | Condensed Matter (Cond-mat), General relativity and quantum cosmology (Gr-qc), High energy physics phenomenology (Hep-ph) e High energy physics theory (Hep-th) | *Dependent Prediction method*, *Weighted Dependent Prediction method*, CN, PA, Katz, *Nonnegative Matrix Factorization* (NMF), *Graph Nonnegative Matrix Factorization* (GNMF) e *Graph Regularized Joint Matrix Factorization* (GRJMF) | Predição de *links* temporal de coautoria |
| [Gao et al. 2012] | Live Journal e arXiv | NMF, *Mixed Membership Stochastic Blockmodels* (MMSB), *Multiplicative Latent Factor Model* (MLFM), *Generalized Latent Factor Model* (GLFM) e *Latent Factor BlockModel* (LFBM) | Predição de *links* de relacionamento social e de coautoria |
| [Gomez Rodriguez and Rogati 2012] | LinkedIn | AA, CN normalizado, Common attendees (CAe) e Adamic-Adar baseado em evento (AAe). | Predição *links* de conexão entre usuários após participarem do mesmo evento |
| [Guo and Guo 2010] | DBLP e TakingltGlobal.org (TIG) | *Merge Weighted Features Algorithm* (MWF) | Predição temporal de *links* de amizades e coautoria baseado em matriz para combinação de características |
| [Hsieh et al. 2013] | LinkedIn, Enron Email e Wiki Talk | CN, AA, *Time overlap*, *Company size*, *Company average age*, *Company cluster coefficient*, *Node propensity e Join time difference* | Predição de *links* interação de usuário aderido há um tempo à rede (com *links*) e de usuário recém-aderido (sem *links*) |

Continua na próxima página.

| Referência | Base de dados | Atributos utilizados | Domínio de aplicação |
|---|---|---|---|
| [Huang et al. 2012] | Epinions | *Average Filling* (AF), JC, SimRank, SVD, *Matrix Completion* (MC) e *Joint Manifold Factorization* (JMF) | Predição de *links* de confiança e desconfiança na rede social através da agregação de redes sociais heterogêneas |
| [Jamali et al. 2011] | Flixster e Epinions | *Generalized Stochastic Blockmodel* (GSBM) e *Mixed Membership Stochastic Blockmodel* (MMB) | Predição de *links* entre usuários em uma Rede Social de Avaliação |
| [Kamei et al. 2012] | @cosme | JC, *Cosine Similarity* (CS) e Modelo probabilístico por-posto com características latentes | Predição de fan-*links* faltantes com base nos dados observados de atividades do usuário |
| [Kunegis et al. 2013] | Epinions e Slashdot | JC, AA, *Exponencial kernel*, PageRank *product*, CN, *Paths of lenght three*, similaridade por cosseno, PA e PageRank condicional | Predição de *links* negativos em rede sociais |
| [Kuo et al. 2013] | Foursquare, Twitter, e DBLP | *User friendship* (UF), *Item ownership* (IO), *Category popularity* (CP), BC, JC, PA, *Attractiveness* (AT), PageRank *with Priors* (PRP), AT-PRP, *Infer* e *Learn* | Predição de *links* unseen-type em uma rede heterogenea |
| [Lerman et al. 2012] | Digg e Twitter | CN, JC, AA, CS, *Attention-limited Conservative Metric* (CS-AL), *Non-Conservative Proximity* (NC) e *Attention-Limited Non-Conservating Proximity* (NC-AL) | Predição de *links* de atividade |
| [Leroy et al. 2010] | Flickr | CN, Katz e *rooted* PageRank (PR) | Predição de *links* entre os usuários em cold start |
| [Liben-Nowell and Kleinberg 2003] | astro-ph, cond-mat, gr-qc, hep-ph e hep-th | CN, JC, AA, PR, Katz, *Hitting time*, SimRankMeta-abordagens: *Low-rank approximation*, *unseen bigrams* e *clustering* | Predição temporal de *links* de co-autoria |
| [Lin et al. 2012] | Interactome, USAir, C. elegance e CGScience | CN, AA, *Resource Allocation* (RA), *Weighted* CN (WCN), *Weighted* AA (WAA), *Weighted Resource Allocation* (WRA), BenefitRanked CN (BrCN), BenefitRanked AA (BrAA) e BenefitRanked RA (BrRA) | Predição de diversos tipos de *links* faltantes em redes ponderadas |
| [Lu et al. 2010] | Hep-th, CiteSeer e Society of Industrial and Applied Mathematics publications (SIAM) | Katz *single source* (Katz-S), Katz *all source* (Katz-C), *Truncated* Katz *single source* ( tKatz-S), *Truncated* Katz *all source* (tKatz-C), *Supervised Learning single source* (SL-S), *Supervised Learning pure color path* (SL-P), SL-P com L1, *Supervised Learning hybrid color paths* (SL-H), SL-H com regularização L1 e SL-H com regularização hierarchical structured (HS) | Predição de *links* de coautoria |
| [Makrehchi 2011] | Informação bibliográfica de publicações em 20 domínios científicos coletados da web | *Latent* Dirichlet *Allocation* (LDA) com Katz, LDA com SP, *Bag-Of-Words* (BOW) e *Latent Semantic Indexing* (LSI) | Predição de *links* de coautoria, a partir da semelhança entre resumos em coautoria entre os autores |
| [Nie et al. 2012] | Wikipedia e Slashdot | CN, SVD, *Fixed Point Continuation* (FPC), *Accelerated Proximal Gradient* (APG) e Método proposto pelos autores | Predição de *links* faltantes de interação entre usuários |
| [Perez et al. 2012] | Um conjunto de redes sociais (Address Book, Twitter, Google+ e Facebook) extraído de iPhones e um conjunto de contatos (amigos e não amigos) extraídos do Facebook | CN, SA, JC, HPI, HDI, LHN, PA, AA, RA e WRA | Predição de *links* para detecção de contatos ilegítimos |
| [Quercia and Capra 2009] | Parte dos dados do projeto Reality Mining do MIT | SP, PageRank, HITS e KmarkovChain | Predição de *links* para recomendar amigos com base na proximidade do celular |
| [de Sa and Prudencio 2011] | DBLP | CN, JC, PA, *Path Distance* (PD), RA, *Local Path* (LP) e *Local Clustering Coefficient* | Predição de *links* temporal de co-autoria em uma rede ponderada |
| [Shin et al. 2012] | Flickr, LiveJournal, MySpace e Epinions | PA, AA, RWR e CN.*Eigen-decomposition* (EIG): EIG-CN e EIG-Katz.*Clustered Low Rank Approximation* (CLRA): CLRA-CN e CLRA-Katz.*Multi-Scale Link Prediction* (MSLP): MSLP-CN e MSLP-Katz | Predição de *links* explorando diferentes escalas de aproximação para redes sociais de grade escala |
| [da Silva Soares and Prudencio 2012] | Hep-th e Hep-lat | AA, PA, CN, JC, *Moving Average* (MA), *Average* (Av), *Random Walk* (RW), *Linear Regression* (LR), *Simple Exponential Smoothing* (SES) e *Linear Exponential Smoothing* (LES) | Predição de *links* de coautoria considerando séries temporais |
| [Song et al. 2009] | Digg, Flickr, LiveJournal, MySpace, YouTube e Wikipedia | PA, PageRank *product* (PRP), CN, AA, Katz, *Graph distance* (GD) | Predição de *links* de relacionamentos em redes sociais de alta escala |
| [Song et al. 2012] | Flickr, LiveJournal e MySpace | Aprendizagem espectral com *Clustered Spectral Graph Embedding* (CSGE), Katz com *Spectral Graph Embedding* (SGE), CN | Predição de *links* e *links* faltantes de relacionamento |

Continua na próxima página.

| Referência | Base de dados | Atributos utilizados | Domínio de aplicação |
|---|---|---|---|
| [Steurer and Trattner 2013] | Second Life (posição dos usuários) e My Second Life | CN, JC, AA, PA, *Physical Distance* (MD), *Common Regions* (RC), *Regions Seen Concurrently* (RS) e *Observations Together* (RO) | Predição de link de interação entre usuários, através de análise dos dados de posição e da rede social |
| [Tian et al. 2010] | Facebook e CALL | Link *trend*, Número de interações totais, Número de recentes interações, Tempo da última interação, número de intervalos de tempo ativo, CN, JC, CN ativos, Número total de amigos, Número total de interações | Predição de *links* para reconexão de *links*. em redes de interação social |
| [Tylenda et al. 2009] | DBLP e astro-ph | Versões de PR e AA padrões e ponderadas por Ano da mais recente colaboração (last), Número de colaborações (count) e Número mínimo de coautores (min. coauth).*Maximum Entropy* (ME), *Time-Aware Maximum Entropy* (TME) *avg., exp.*, TME *avg. lin.*, TME *avg. sqrt.*, TME *sum lin.*, Distance (dist), JC, CN, *last count*, *count last*, *min. coauthors*, *dist. last count*, *dist. count last*, *dist. min. coauth*, ordenação por *count last*, ordenação por *last count* | Predição de *links* de coautoria, novos e repetidos |
| [Valverde-Rebaza and Lopes 2012] | Twitter | *Within And Inter Cluster* (WIC), CN, AA, JC, RA e PA | Predição de *links* de seguidores no Twitter |
| [Vasuki et al. 2010] | Orkut e Youtube | tkatz e SVD | Predição de *links* para recomendação de comunidades |
| [Vasuki et al. 2011] | Rede combinada com componentes altamente conectados do Orkut (Orkut-Icc) e Youtube (Youtube-lcc) | tKatz-C, tKatz-A, tKatz com *latent factor model* (tKatz-LFM), tKatz com *common subspace model* (tKatz-CS) e tKatz *clustered latent factor model Equation* (tKatz-LFM-c) | Predição de *links* para recomendação de grupos ou comunidades em redes de grande escala |
| [Wang et al. 2011A] | CDRs | Katz, AA, CN, JC, *Spatial Cosine Similarity*, *Weighted Spatial Cosine Similarity*, *Extra-role Co-Location Rate Weighted*, *Weighted Co-Location Rate Common* e *Co-Location Rate* | Predição de *links* de chamadas com medidas de mobilidade |
| [Wang et al. 2011b] | CORA | *Dynamic Relational Topic Model* (dRTM) e RTM | Predição de *links* de citação com um modelo capaz de lidar com *links* ruidosos |
| [Wang et al. 2007] | DBLP, Genetics e Biochemistry | *Approximate Katz measure* (aKatz), *Co-occurrence probability*, AA e PA | Predição de *links* de coautoria utilizando um novo modelo probabilístico em rede de coautoria |
| [Xia et al. 2012] | Internet Movie Database | CN, JC, AA, *Collaborative Filtering* (CF), PA, Katz, *Minimum Description Length* (MDL), *Absent Links* (AL) e *Random Walk with Restart* (RWR) | Predição de *links* entre diretor e ator de filmes com métodos adaptados de métodos tradicionais baseados em vizinhança para redes sociais bipartidas |
| [Yin et al. 2011] | Twitter | Método proposto, PropFlow, Katz, JC, AA, CN, PA e *Matrix factorization* | Predição de *links* de seguidores em uma rede híbrida |
| [Yu et al. 2011] | Geraram quatro conjuntos de dados sintéticos e o conjunto de dados MIT Reality Mining Project | *Random*, *Same Edge*, GPS *Similarity*, RWR e *Geo-Friends Recommendation Framework* (GEFR) | Predição de *links* para recomendação de amigos em uma rede social cyber-physical |
| [Zhang et al. 2013] | Sina Microblog | *Exponential random graph model* (ERGM), JC e Katz | Predição de *links* de relacionamento nas comunidades de um microblog |
| [Zhong et al. 2013] | Tencent, SinaWeibo, Xiaone, Facebook, Twitter, Github, Stackoverflow e Epinions | *Time-evolving Composite Network Models* (ITCom), *Mixed Membership Stochastic Blockmodels* (MMSB), *dynamic Mixed Membership Stochastic Blockmodels* (dMMSB), *Nonparametric Metadata Dependent Relational Model* (NMDR), *dynamic Infinite Relational Model* (dIRM) e *Tensor Factorization* (TF) | Predição temporal de *links* de interação e amizade entre usuários |

Com o intuito de predizer novos *links* considerando o comportamento dos *links* conforme a série temporal, [da Silva Soares and Bastos Cavalcante Prudencio 2012] utilizaram duas sessões do arXiv: *Theoretical high energy physics area* (1991 a 2010) e *High energy physics - lattice area* (1993 a 2010). A ideia básica é a construção de séries temporais para cada par de nós não conectados, usando um *score* de similaridade calculado por uma métrica topológica. Um modelo de previsão é então utilizado, a fim de prever o valor seguinte da série. Esse valor é o resultado final do par de nós para ser usado pelos métodos de predição de *link*, testado conforme uma abordagem supervisionada e não supervisionada. Conforme os autores, a abordagem supervisionada foi

melhor em todos os modelos de previsão em relação à abordagem não supervisionada, mas este trabalho ainda apresenta limitações quanto ao número de redes utilizadas nos experimentos e seus domínios.

Em [de Sa and Prudencio 2011] os autores trabalharam com redes acadêmicas e investigaram a relevância do uso de pesos nas ligações (arestas). Os autores utilizaram o conjunto de dados do DBLP, dividido em três conjuntos: não ponderada caso os dois autores já foram coautores de um mesmo artigo, ponderada de acordo com o número total de trabalhos em que o par de autores foi coautores e ponderada pela contribuição dos autores em seus trabalhos de coautoria. Conforme mostram os autores, em quase todas as comparações entre as redes, a rede não ponderada obteve um desempenho inferior em relação a, pelo menos, uma das redes ponderadas. Os autores concluem que, embora estes resultados não sejam conclusivos, é possível realizar melhorias no desempenho da predição de *links* ao se considerar os pesos de cada ligação.

Algumas pesquisas como a de [Gomez Rodriguez and Rogati 2012] são mais complexas ao considerar não só a interação *online*, mas também a interação *offline*, através de encontros sociais ou profissionais, entre os usuários. Com o intuito de mostrar como os eventos profissionais e encontros sociais no mundo real se relacionam com a dinâmica temporal e evolução de uma rede profissional, os autores concluíram que novos *links* são realizados em curto período após a data do evento e que sua predição é mais eficiente nesse período. Além disso, a conexão entre nós distintos possui influência dos nós em comum que ambos compartilham.

Em [Wang et al. 2011a], o foco principal do trabalho é explorar o poder preditivo de mobilidade individual comparado e combinado com atributos topológicos. Para tal, utilizaram as trajetórias e os padrões de comunicação de uma base anônima de um país, cujos dados são obtidos de CDRs (*Call Detail Record*). Segundo os autores, os resultados demonstram que a mobilidade tem forte influência na predição de *links*, conforme a correlação entre a semelhança entre os movimentos dos indivíduos, suas conexões sociais e a força das interações entre eles. Combinando as medidas de mobilidade e de rede, os autores mostraram que a precisão na predição pode ser significativamente melhorada com aprendizado supervisionado.

## 4. Considerações Finais

Com a revisão realizada é visível que o tema predição de *links* para redes sociais é recente e ao analisar todos os trabalhos revisados como um conjunto, pode-se inferir que há alguns atributos considerados tradicionais (ou de referência) como CN, Katz, JC, AA e PA, já que podem ser aplicados em diferentes domínios. Além disso, são utilizados como base para comparar o desempenho de métodos propostos ou são utilizados como base para criar variações.

O sistema de estabelecimento de relacionamentos é muito complexo, além da dinamicidade, às vezes engloba diversos fatores (como acontecimentos externos) e a alta escala na maioria das redes. Além disso, foi possível entender que cada rede social possui suas características, logo não há atributos ideais que satisfaçam todas as redes sociais. Com isso, nota-se uma grande variedade de atributos, a qual pode ser verificado na Tabela 1.

Deste modo, pode-se concluir, com base no levantamento realizado no presente trabalho, que o tema tem ganho destaque nos últimos anos e, sendo um tema recente, há muito a se pesquisar sobre novos atributos e domínios de aplicação para entendimento da complexidade de uma rede social. Mesmo assim, o presente trabalho pode ser utilizado como leitura básica para os pesquisadores ou desenvolvedores que objetivem criar novos sistemas de predição em uma rede social específica ou um algoritmo de predição multi-domínio.

## Acknowledgments

O trabalho apresentado neste artigo foi parcialmente financiado pela CAPES e pelo CNPq (processos 306046/2013-0 e 477246/2013-3).

## Referências

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., and Menczer, F. (2012). Friendship prediction and homophily in social media. *ACM Trans. Web*, 6(2):9:1–9:33.

Almansoori, W., Gao, S., Jarada, T., Alhajj, R., and Rokne, J. (2011). Link prediction and classification in social networks and its application in healthcare. In *Information Reuse and Integration (IRI), 2011 IEEE International Conference on*, pages 422–428.

Biolchini, J., Mian, P. G., Candida, A., and Natali, C. (2005). Systematic Review in Software Engineering. Technical Report May, Systems Engineering and Computer Science Department, COPPE/UFRJ, Rio de Janeiro.

Chang, C. and Yao, X. (2011). Social network link predict based on af model. In *Computer Science and Network Technology (ICCSNT), 2011 International Conference on*, volume 1, pages 415–418.

Chelmis, C. and Prasanna, V. (2012). Predicting communication intention in social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 184–194.

Corlette, D. and Shipman, III, F. M. (2010). Link prediction applied to an open large-scale online social network. In *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*, HT '10, pages 135–140, New York, NY, USA. ACM.

Costa, G. and Ortale, R. (2012). A bayesian hierarchical approach for exploratory analysis of communities and roles in social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 194–201.

Cukierski, W., Hamner, B., and Yang, B. (2011). Graph-based features for supervised link prediction. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1237–1244.

da Silva Soares, P. and Bastos Cavalcante Prudencio, R. (2012). Time series based link prediction. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–7.

de Sa, H. and Prudencio, R. (2011). Supervised link prediction in weighted networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2281–2288.

Dhote, Y., Mishra, N., and Sharma, S. (2013). Survey and analysis of temporal link prediction in online social networks. In *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*, pages 1178–1183.

Digiampietri, L. A., do Nascimento Santiago, C. R., and Alves, C. M. (2013). Predição de coautorias em redes sociais acadêmicas: um estudo exploratório em Ciência da Computação. In *II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2013)*, page 12, Maceió, Alagoas, Brasil.

Digiampietri, L. A. and Maruyama, W. T. (2014). Predição de novas coautorias na rede social acadêmica dos programas brasileiros de pós-graduação em ciência da computação. In *III Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2014)*, pages 243–248.

Digiampietri, L. A., Maruyama, W. T., Santiago, C. R. N., and da Silva Lima, J. J. (2015). Um sistema de predição de relacionamentos em redes sociais. In *XI Simpósio Brasileiro de Sistemas de Informação (SBSI 2015)*, pages 139–146.

Dong, Y., Ke, Q., Rao, J., Wang, B., and Wu, B. (2011a). Random walk based resource allocation: Predicting and recommending links in cross-operator mobile communication networks. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 358–365.

Dong, Y., Ke, Q., Rao, J., and Wu, B. (2011b). Predicting missing links via local feature of common neighbors. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference on*, volume 2, pages 1038–1042.

Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N., Rao, J., and Cao, H. (2012). Link prediction and recommendation across heterogeneous social networks. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 181–190.

Fire, M., Tenenboim, L., Lesser, O., Puzis, R., Rokach, L., and Elovici, Y. (2011). Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80.

Gao, S., Denoyer, L., and Gallinari, P. (2011). Temporal link prediction by integrating content and structure information. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1169–1174, New York, NY, USA. ACM.

Gao, S., Denoyer, L., and Gallinari, P. (2012). Link prediction via latent factor blockmodel. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 507–508, New York, NY, USA. ACM.

Gomez Rodriguez, M. and Rogati, M. (2012). Bridging offline and online social graph dynamics. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2447–2450, New York, NY, USA. ACM.

Guo, J. and Guo, H. (2010). Multi-features link prediction based on matrix. In *Computer Design and Applications (ICCDA), 2010 International Conference on*, volume 1, pages V1–357–V1–361.

Hasan, M. and Zaki, M. (2011). A survey of link prediction in social networks. In Aggarwal, C. C., editor, *Social Network Data Analytics*, pages 243–275. Springer US.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

Hsieh, C.-J., Tiwari, M., Agarwal, D., Huang, X. L., and Shah, S. (2013). Organizational overlap on social networks and its applications. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 571–582, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Huang, J., Nie, F., Huang, H., and Tu, Y.-C. (2012). Trust prediction via aggregating heterogeneous social networks. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1774–1778, New York, NY, USA. ACM.

Jamali, M., Huang, T., and Ester, M. (2011). A generalized stochastic block model for recommendation in social rating networks. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 53–60, New York, NY, USA. ACM.

Kamei, T., Ono, K., Kumano, M., and Kimura, M. (2012). Predicting missing links in social networks with hierarchical dirichlet processes. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8.

Kunegis, J., Preusse, J., and Schwagereit, F. (2013). What is the added value of negative links in online social networks? In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 727–736, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Kuo, T.-T., Yan, R., Huang, Y.-Y., Kung, P.-H., and Lin, S.-D. (2013). Unsupervised link prediction using aggregative statistics on heterogeneous social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 775–783, New York, NY, USA. ACM.

Lerman, K., Intagorn, S., Kang, J.-H., and Ghosh, R. (2012). Using proximity to predict activity in social networks. In *Proceedings of the 21st International Conference Companion on World Wide Web*, WWW '12 Companion, pages 555–556, New York, NY, USA. ACM.

Leroy, V., Cambazoglu, B. B., and Bonchi, F. (2010). Cold start link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 393–402, New York, NY, USA. ACM.

Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, CIKM '03, pages 556–559, New York, NY, USA. ACM.

Lin, Z., Yun, X., and Zhu, Y. (2012). Link prediction using benefitranks in weighted networks. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 423–430, Washington, DC, USA. IEEE Computer Society.

Lu, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150 – 1170.

Lu, Z., Savas, B., Tang, W., and Dhillon, I. (2010). Supervised link prediction using multiple sources. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 923–928.

Makrehchi, M. (2011). Social link recommendation by learning hidden topics. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 189–196, New York, NY, USA. ACM.

Maruyama, W. and Digiampietri, L. (2016). Co-authorship prediction in academic social network. In *BRASNAM 2016*.

Murata, T. and Moriyasu, S. (2008). Link prediction based on structural properties of online social networks. *New Generation Computing*, 26(3):245–257.

Nie, F., Wang, H., Cai, X., Huang, H., and Ding, C. (2012). Robust matrix completion via joint schatten p-norm and lp-norm minimization. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 566–574.

Perez, C., Birregah, B., and Lemercier, M. (2012). The multi-layer imbrication for data leakage prevention from mobile devices. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*, pages 813–819.

Quercia, D. and Capra, L. (2009). Friendsensing: Recommending friends using mobile phones. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, pages 273–276, New York, NY, USA. ACM.

Shin, D., Si, S., and Dhillon, I. S. (2012). Multi-scale link prediction. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 215–224, New York, NY, USA. ACM.

Song, H. H., Cho, T. W., Dave, V., Zhang, Y., and Qiu, L. (2009). Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, IMC '09, pages 322–335, New York, NY, USA. ACM.

Song, H. H., Savas, B., Cho, T. W., Dave, V., Lu, Z., Dhillon, I. S., Zhang, Y., and Qiu, L. (2012). Clustered embedding of massive social networks. *SIGMETRICS Perform. Eval. Rev.*, 40(1):331–342.

Steurer, M. and Trattner, C. (2013). Predicting interactions in online social networks: An experiment in second life. In *Proceedings of the 4th International Workshop on Modeling Social Media*, MSM '13, pages 5:1–5:8, New York, NY, USA. ACM.

Tian, Y., He, Q., Zhao, Q., Liu, X., and Lee, W.-c. (2010). Boosting social network connectivity with link revival. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 589–598, New York, NY, USA. ACM.

Tylenda, T., Angelova, R., and Bedathur, S. (2009). Towards time-aware link prediction in evolving social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, SNA-KDD '09, pages 9:1–9:10, New York, NY, USA. ACM.

Valverde-Rebaza, J. and de Andrade Lopes, A. (2012). Structural link prediction using community information on twitter. In *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, pages 132–137.

Vasuki, V., Natarajan, N., Lu, Z., and Dhillon, I. S. (2010). Affiliation recommendation using auxiliary networks. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 103–110, New York, NY, USA. ACM.

Vasuki, V., Natarajan, N., Lu, Z., Savas, B., and Dhillon, I. (2011). Scalable affiliation recommendation using auxiliary networks. *ACM Trans. Intell. Syst. Technol.*, 3(1):3:1–3:20.

Wang, C., Satuluri, V., and Parthasarathy, S. (2007). Local probabilistic models for link prediction. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 322–331.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., and Barabasi, A.-L. (2011a). Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1100–1108, New York, NY, USA. ACM.

Wang, E., Silva, J., Willett, R., and Carin, L. (2011b). Dynamic relational topic model for social network analysis with noisy links. In *Statistical Signal Processing Workshop (SSP), 2011 IEEE*, pages 497–500.

Xia, S., Dai, B., Lim, E.-P., Zhang, Y., and Xing, C. (2012). Link prediction for bipartite social networks: The role of structural holes. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 153–157.

Yin, D., Hong, L., and Davison, B. D. (2011). Structural link analysis and prediction in microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1163–1168, New York, NY, USA. ACM.

Yu, X., Pan, A., Tang, L.-A., Li, Z., and Han, J. (2011). Geo-friends recommendation in gps-based cyber-physical social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 361–368.

Zhang, C., Zhai, B. Y., and Wu, M. (2013). Link prediction of community in microblog based on exponential random graph model. In *Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on*, pages 1–6.

Zhong, E., Fan, W., Zhu, Y., and Yang, Q. (2013). Modeling the dynamics of composite social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 937–945, New York, NY, USA. ACM.