UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
http://www.icmc.usp.br/

Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas

Proposta de Projeto de Mestrado

Orientador: Prof. Dr. Diego Raphael Amancio[1]
Candidata: Vanessa Queiroz Marinho[2]
diego@icmc.usp.br[1], vanessaqm@usp.br[2]

Universidade de São Paulo (USP)
Instituto de Ciências Matemáticas e de Computação (ICMC)
Avenida Trabalhador são-carlense, 400 - Centro
CEP: 13566-590 - São Carlos - SP

29 de março de 2015

Desenvolvimento de novos modelos para reconhecimento de autoria com a utilização de redes complexas

Instituto de Ciências Matemáticas e de Computação (ICMC)
Universidade de São Paulo (USP), São Carlos/SP, Brasil

Orientador: Prof. Dr. Diego Raphael Amancio
Candidata: Vanessa Queiroz Marinho

Resumo

A modelagem de grafos e redes complexas vem sendo aplicada com sucesso em diferentes domínios, sendo objeto de estudo de distintas áreas que incluem, por exemplo, a matemática e a computação. A descoberta de que métodos derivados do estudo de redes complexas podem ser utilizados para analisar textos em seus distintos níveis de complexidade proporcionou grandes avanços em tarefas de processamento de línguas naturais. Exemplos de aplicações analisadas com os métodos e ferramentas de redes complexas são a detecção de conceitos relevantes, a criação de sumarizadores extrativos automáticos e reconhecedores de autoria. Esta última tarefa, que é foco deste projeto de pesquisa, tem sido estudada com certo sucesso através da representação de redes de adjacência de palavras que conectam apenas as palavras mais próximas. O objetivo deste projeto é estender a modelagem tradicional, escolhendo-se a janela de conexão ótima para o problema, para um dado conjunto de treinamento. Além disso, pretende-se utilizar informação de conectividade de palavras funcionais para complementar a caracterização de estilo de autores. Finalmente, pretende-se criar classificadores híbridos que sejam capazes de combinar fatores tradicionais com as propriedades fornecidas pela análise topológica de redes complexas. Através da adaptação, combinação e aperfeiçoamento da modelagem, pretendemos não apenas melhorar o desempenho dos sistemas de caracterização estilística textual e reconhecimento de autoria, mas também entender melhor quais são os fatores quantitativos textuais (medidos via redes) que podem ser utilizados na área de estilometria. Os avanços obtidos durante este projeto podem ser úteis para estudar aplicações relacionadas, como é o caso de análise de inconstistências estilísticas e plagiarismos.

1 Introdução

A modelagem de sistemas reais por meio de redes complexas têm sido útil para descrever uma grande variedade de sistemas encontrados na natureza e na sociedade [1]. Alguns exemplos incluem a célula, que pode ser descrita como uma rede de substâncias conectadas por reações químicas, e a Internet, uma rede de roteadores e computadores conectados por links físicos [2]. Muitos trabalhos científicos na área de redes complexas são interdisciplinares, beneficiando-se de ideias provenientes de diferentes disciplinas, como matemática, física, biologia, ciência da computação, ciências sociais e outras [3]. Tradicionalmente, o estudo de redes limitava-se à teoria dos grafos aplicada a sistemas aleatórios. Nos últimos anos, no entanto, a descoberta de que vários sistemas reais podem ser caracterizados por uma estrutura de rede não aleatória [3] permitiu um rápido desenvolvimento da área. A informatização e disponibilização de dados de várias áreas do conhecimento e o aumento do poder computacional propiciaram a análise e o desenvolvimento de algoritmos eficientes em uma ampla gama de aplicações de naturezas distintas [4]. De particular interesse para os objetivos deste projeto, podemos citar as redes textuais, que podem ser formadas por relações sintáticas [5], semânticas [6] ou empíricas [7]. Nestes três tipos de redes, as propriedades universais *small-world* e *scale-free* foram encontradas [8].

Um caso particular das redes sintáticas são as redes de co-ocorrência de palavras (ou redes de adjacência de palavras) [8]. Em tais redes, as conexões sintáticas são aproximadas pelas conexões de palavras adjacentes, já que a maioria das conexões sintáticas ocorre entre palavras vizinhas [9]. O sucesso da representação de elementos textuais como redes de ocorrências pode ser comprovado a partir de aplicações de reconhecimento de movimentos literários [10], sumarização automática [5], entre outras. Um estudo das propriedades de tais redes desenvolvido em [11] demonstrou que a maioria das medidas topológicas da representação por redes captura características da sintaxe e estilo da linguagem. Por este motivo, estas redes são mais adequadas para tratar problemas de estilo, embora uma dependência entre topologia e semântica de palavras tenha sido também comprovada [12]. Neste contexto, pretendemos fazer uso do poder discriminativo gerado por redes de co-ocorrência de palavras para tratar discriminar estilos na tarefa de reconhecimento de autoria.

Métodos de reconhecimento de autoria são relevantes pois estes podem ser aplicados para classificar obras literárias, resolver direitos autorais [13] ou até mesmo para interceptar mensagens terroristas [14]. As primeiras técnicas de reconhecimento de autoria foram exploradas após o famoso trabalho de Mosteller e Wallace na análise dos artigos federalistas [15]. Após este estudo, pesquisadores têm proposto novos atributos para caracterizar estilos [16]. Alguns dos atributos tradicionalmente utilizados para a tarefa incluem a análise das propriedades estatísticas de palavras (p.e. o comprimento médio, a frequência, o *burstiness* e a riqueza de vocabulário) [4] e caracteres (frequências e correlações) [13]. Além dos atributos léxicos, atributos sintáticos (p.e. frequências de *chunks* específicos) e semânticos também têm sido utilizados como atributos úteis para o problema [4]. Atualmente, novos atributos têm sido propostos para geração de classificadores robustos [4]. A adequabilidade das medidas de redes complexas para a tarefa foi observada pela primeira vez em [11], onde foi mostrado que

as características topológicas das redes podem ser utilizadas para identificar autores. Este projeto pretende estender este modelo com a introdução de novas modelagens e novas formas de caracterizar estilos. Além disso, pretendemos analisar a robustez da modelagem com relação ao tamanho do texto, escolha de palavras específicas e outros fatores. Como a análise topológica não considera os rótulos dos vértices, pretendemos incluir essa informação através da implementação de classificadores híbridos. Desse modo, duas componentes serão consideradas, a componente topológica (medidas de redes complexas) e a componente proveniente de métodos tradicionais, como a frequência de palavras funcionais específicas.

De maneira geral, este projeto pretende estender o trabalho realizado em projetos anteriores financiados pela FAPESP (processos 10/00927-9 e 13/06717-4) ao combinar técnicas tradicionais com técnicas de redes. Espera-se, portanto, uma melhoria de performance da tarefa, já que as desvantagens de cada técnica devem ser superadas pela classificação híbrida. Além de contribuir para melhorar o desempenho da tarefa, esperamos criar classificadores mais robustos que os atuais, que são vulneráveis a ataques [17]. Durante o desenvolvimento deste projeto, a aluna pretende realizar estágio de pesquisa no exterior com duração máxima de 6 meses. Acredita-se que esta será uma etapa importante para o desenvolvimento deste projeto de mestrado, mais informações sobre o estágio encontram-se na Seção 7. Os detalhes desta proposta são apresentados nas próximas seções. Na Seção 2, a motivação e os objetivos desta proposta são apresentados. A definição de redes complexas, sua utilização na modelagem de textos e algumas medidas aplicadas a essas redes são apresentadas na Seção 3. A Seção 4 apresenta diversos trabalhos e aplicações de reconhecimento de autoria. Na Seção 5, algumas técnicas de reconhecimento de padrões são descritas. Na Seção 6, as atividades propostas e os resultados esperados são apresentados. Por fim, na Seção 7, encontra-se o plano de atividades a ser seguido na execução do projeto.

2 Motivação e objetivos gerais

Desde o fim da década de 90, os estudos de atribuição de autoria passaram por grandes mudanças. Com a popularização e facilidade de acesso à rede mundial de computadores, uma grande quantidade de textos eletrônicos tornou-se disponível na Internet (por exemplo e-mails, posts em blogs e fóruns) e aumentou a necessidade por métodos eficientes de tratamento dessas informações. Devido ao aumento do poder computacional, os métodos de atribuição de autoria, que antes eram assistidos por computador, passaram a ser baseados em computador, com o desenvolvimento de sistemas totalmente automatizados [4]. Essas mudanças trouxeram um grande e positivo impacto em áreas como aprendizado de máquina e processamento de línguas naturais.

Nos últimos anos, a quantidade de textos disponíveis e de fácil acesso na Web revelou o potencial da análise de autoria em diferentes aplicações de diversas áreas. Esta tarefa é bastante relevante dentro da área de processamento de línguas naturais contribuindo para diversos avanços na literatura [18], história [19], serviços de inteligência [14], computação forense [16, 20] e também em investigações criminais [21]. Outra aplicação importante desta tarefa se dá no contexto do plagiarismo, pois é possível identificar trechos de plágios e incon-

sistências estilísticas a partir da tarefa de verificação de autoria. A importância desta tarefa se torna ainda mais evidente também quando se deseja estimar a similaridade de textos [22].

Embora redes complexas já tenham sido utilizadas para reconhecer autoria [11], o estado da arte ainda não foi atingido [11]. Por este motivo, o objetivo deste projeto de mestrado é tentar aperfeiçoar os atuais modelos baseados em redes através da extensão do modelo atual desenvolvido em [11] e concepção/uso de novas medidas de análise que sejam mais dependentes do estilo. Uma vez que os modelos de rede são recentes, estes nunca foram aplicados em combinação com técnicas tradicionais. Dessa forma, este projeto visa também estudar métodos de combinação de classificadores híbridos que forneçam a combinação mais adequada para o problema. Além da contribuição esperada em termos de desempenho, os métodos de redes serão adaptados para situações em que exista a possibilidade de ataque (autores disfarçando o estilo) e em casos onde os textos obtidos sejam curtos. Os avanços obtidos neste projeto devem ser úteis não apenas para o reconhecimento de autoria, mas também para identificar plagiarismos caracterizados por textos com estrutura similar.

3 Redes complexas

3.1 Redes complexas e linguagens

Nos últimos anos, a teoria dos grafos tem sido largamente utilizada para provar propriedades da linguagem. Redes/grafos podem ser utilizados para modelar objetos e as conexões existentes entre eles. Um grafo $G = (V,A)$ é uma estrutura matemática que consiste de dois conjuntos finitos V e A, em que os elementos de V são denominados vértices (ou nós), e os elementos de A são denominados arestas. Cada aresta conecta dois vértices da rede [23]. Este tipo de modelagem é aplicado na representação de dados provenientes das mais diversas áreas científicas, incluindo a matemática discreta [24] e computação [25].

Diferentemente da análise tradicional de grafos, as redes complexas modelam sistemas usualmente envolvendo a interação de uma grande quantidade de indivíduos de maneira não trivial [3]. Através da representação das relações entre entidades como arestas conectando vértices, sistemas reais como as redes de transporte, a Internet, a *World Wide Web*, os sistemas biológicos e as redes sociais puderam ser analisados com a teoria de sistemas complexos. O uso deste arcabouço de análise permitiu inferir diversos padrões que não seriam revelados com modelagens tradicionais baseadas em análises frequencionistas e em redes aleatórias de pequena escala [26]. Dentre os principais padrões encontrados destaca-se a emergência da propriedade livre de escala [27], caracterizada pela distribuição de probabilidade $P(k)$ do número de conexões k de acordo com a lei

$$P(k) \simeq k^{-\gamma}, \tag{1}$$

onde γ geralmente está no intervalo $2 < \gamma < 3$. Outro padrão emergindo dos sistemas complexos reais é conhecido como fenômeno de pequeno mundo [28], o qual afirma que a distância típica L entre duas entidades varia com o logaritmo do tamanho M da rede, isto é, $L \propto \log M$. Ainda merecem destaque a organização topológica em comunidades [29]

(i.e., grupos de vértices fortemente conectados) e a assortativade [30], que corresponde à tendência de conexões preferenciais entre vértices compartilhando propriedades similares. Na modelagem de sistemas reais como redes complexas, destacam-se estudos teóricos para estabelecer condições de propagação e controle de doenças em redes sociais [31, 32] e a análise de redes neuronais.

Até o final da década de 90, acreditava-se que a maioria dos sistemas reais podiam ser representados por grafos aleatórios. A introdução dos modelos *scale-free* e *small-world* mostrou que muitos sistemas, na verdade, apresentam padrões heterogêneos de conectividade [3]. Interessantemente, neste período, foi mostrado que a linguagem apresenta padrões específicos de conectividade semelhantes a outros sistemas reais [8]. A partir de tais descobertas, muitos problemas linguísticos passaram a ser analisados com redes complexas. Diversas aplicações de redes para análise de problemas linguísticos são descritas em [2], como sumarizadores de texto, identificadores de tópicos, além de aplicações em tarefas de tradução automática e processamento de fala. Um modelo que tem sido bastante utilizado na literatura é o modelo de rede de adjacência de palavras (também conhecido como rede de co-ocorrência), que realiza a conexão de palavras próximas. Uma vez que este modelo consegue representar fatores sintáticos e estilísticos [22], ele tem sido empregado em análises de complexidade sintática [33], detecção de movimentos literários[10] e em estilometria [34]. Devido ao sucesso deste modelo em capturar características estilísticas de textos, tal modelo será utilizado como base para o desenvolvimento deste projeto. Nas Seçoes 3.2 e 3.3, respectivamente, mostramos como criar redes de textos e como analisá-las topologicamente.

3.2 Modelando textos como redes complexas

A modelagem de textos como redes complexas pode ser dividida em duas etapas, o pré-processamento do texto e a conexão de conceitos. Uma das atividades iniciais de pré-processamento é a retirada de sinais de pontuação. Um procedimento tipicamente adotado em trabalhos de literatura [10] é a extração das *stopwords* ou *function words*, por apresentarem pouco conteúdo semântico. Porém, durante o desenvolvimento deste trabalho, as *stopwords* serão incluídas na construção da rede de palavras. O objetivo é analisar o aumento do poder descritivo da rede para a tarefa de reconhecimento de autoria, visto que tais palavras são úteis em estratégias tradicionais de reconhecimento de autoria não baseadas em grafos. Outra etapa do pré-processamento é a lematização. Essa etapa é aplicada nas palavras remanescentes com a utilização de um rotulador de classes gramaticais (*part-of-speech tagger*). Palavras no plural são transformadas para sua forma singular, verbos são transformados para sua forma infinitiva e nomes são convertidos para sua forma masculina. Assim, palavras referentes a um mesmo conceito serão associadas a um mesmo vértice, independente das variações de flexão. O rotulador de classes gramaticais utilizado para este exemplo foi o NLTK (*Natural Language Toolkit*) [35]. Para exemplificar os passos de pré-processamento que serão utilizados neste trabalho, a Tabela 1 ilustra suas aplicações em um extrato em inglês do livro *"Alice's Adventures in Wonderland"* ("Alice no País das Maravilhas"), de Lewis Carroll.

Após o pré-processamento, é necessário realizar a conexão de conceitos. Por ser formada

Tabela 1: Exemplo de aplicação do pré-processamento para modelagem de textos como redes. Um extrato obtido do livro *"Alice's Adventures in Wonderland"* ("Alice no País das Maravilhas"), de Lewis Carroll está ilustrado após a remoção dos sinais de pontuação e subsequente lematização

Texto Original	**Após pré-processamento**
Can you play croquet? shouted the Queen. The question was evidently meant for Alice. "Yes!" said Alice loudly.	*can you play croquet shout the queen the question be evidently mean for alice yes say alice loudly*

por cadeias lineares de palavras, o modo mais simples de representar a linguagem escrita é conectar palavras adjacentes. Este tipo de rede, conhecido como rede de co-ocorrência, é amplamente utilizado na literatura [8, 11, 36]. Como a maioria das relações sintáticas acontece na primeira vizinhança, esta modelagem pode ser vista como uma aproximação das ligações sintáticas [9]. Em uma rede de co-ocorrência, os vértices representam palavras e as arestas são estabelecidas entre palavras vizinhas. A rede obtida com o exemplo da Tabela 1 está ilustrada na Figura 1. Esta rede foi criada conectando-se o primeiro vizinho mais próximo de cada palavra. Outro fator importante na modelagem de co-ocorrência refere-se à

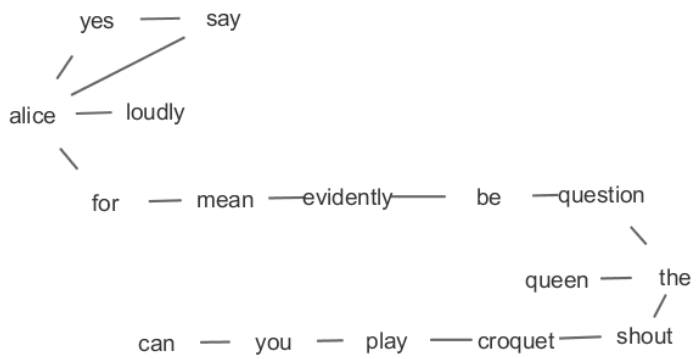


Figura 1: Subgrafo não orientado obtido para a sentença *"Can you play croquet? shouted the Queen. The question was evidently meant for Alice. Yes! said Alice loudly."* obtida do livro "Alice's Adventures of Wonderland", de Lewis Carroll.

escolha do tamanho da janela de conexão de palavras. A abordagem de conexão de palavras imediatamente vizinhas, como na Figura 1, representa uma simplificação considerável, já que correlações de longo alcance estão presentes na linguagem humana [37]. Neste trabalho, iremos analisar a efetividade da modelagem para diferentes janelas de conexão. Além disso, em uma modelagem diferente, todas as palavras de uma sentença serão conectadas. O objetivo é que, com essas alterações na modelagem, seja possível capturar uma quantidade maior de links relevantes entre palavras e caracterizar melhor o contexto. A Figura 2 ilustra o subgrafo construído, para a mesma sentença do livro *"Alice's Adventures in Wonderland"*, utilizando

uma janela de conexão de tamanho 2.

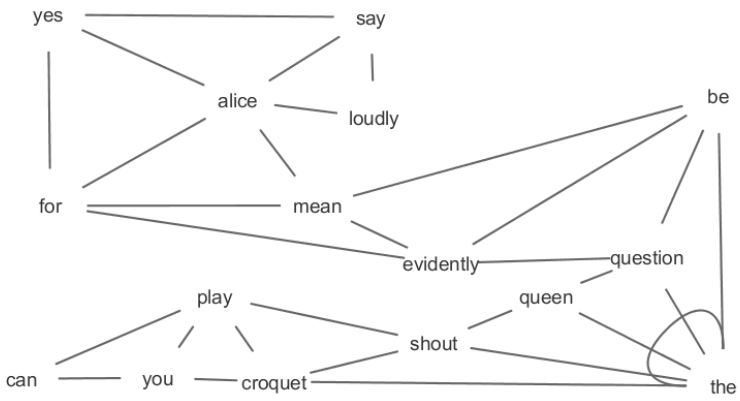


Figura 2: Subgrafo obtido utilizando uma janela de conexão de tamanho 2.

Vários são os exemplos da utilidade das redes de co-ocorrência para modelagem de textos [10, 5, 38, 39]. É importante ressaltar que a representação por redes e os métodos de classificação são genéricos, sendo aplicáveis em várias tarefas de processamento de línguas naturais, não apenas em reconhecimento de autoria.

3.3 Medidas de redes complexas

As medidas que serão utilizadas para caracterizar a estrutura de textos são divididas entre medidas de intermitência e medidas topológicas de redes. Embora o primeiro tipo de medida possa ser considerado uma medida topológica de rede, por ter relação com o número de ciclos na rede, ela é calculada diretamente do texto pré-processado. [22]. Entre as medidas topológicas mais relevantes de redes complexas, e que serão utilizadas neste trabalho, podem-se citar:

- **Assortatividade**: muitas redes reais apresentam correlações de grau, isto é, a tendência de vértices se conectarem com outros vértices de conectividade similar. Um modo de se medir correlações é através do coeficiente de correlação de Pearson r do grau $k_i = \sum_j a_{ij}$ de todos os pares de vértices [3]:

$$r = \frac{e^{-1}\sum_{j>i} k_i k_j a_{ij} - \left[e^{-1}\sum_{j>i} \frac{1}{2}(k_i + k_j)a_{ij}\right]^2}{e^{-1}\sum_{j>i} \frac{1}{2}(k_i^2 + k_j^2)a_{ij} - \left[e^{-1}\sum_{j>i} \frac{1}{2}(k_i + k_j)a_{ij}\right]^2}, \tag{2}$$

  onde $e = \frac{1}{2}\sum_i k_i$ representa o número de arestas e a_{ij} é o elemento da matriz de adjacência que indica se os vértices i e j estão conectados. Em redes assortativas (com r > 0), vértices com alto grau tendem a se conectar com outros vértices de alto grau. Em contrapartida, em redes disassortativas (r < 0), vértices de baixo grau são conectados com vértices de baixo grau. Na maioria dos casos, as redes de adjacência de palavras são disassortativas.

- **Aglomeração** (*Clustering Coefficient*): o coeficiente de aglomeração, C_i, é dado pelo número real de conexões entre vizinhos de um vértice i dividido pelo máximo número possível de conexões entre eles. Portanto, esta medida quantifica a densidade de conexões entre vizinhos, indicando a probabilidade de que dois vizinhos de um vértice estejam conectados entre si. Esta medida é calculada como:

$$C_i = \frac{3 \sum_{k>j>i} a_{ij} a_{ik} a_{jk}}{\sum_{k>j>i} a_{ij} a_{ik} + a_{ji} a_{jk} + a_{ki} a_{kj}} \tag{3}$$

  Sobre essa medida, os autores de [8] constataram que o coeficiente de aglomeração em redes representando textos é muito maior que o valor esperado para a correspondente rede aleatória de palavras.

- **Caminhos mínimos**: o comprimento médio dos caminhos mínimos, L, representa o número médio de passos necessários para alcançar qualquer vértice da rede, considerando-se todos os pares de vértices. Em redes textuais, L quantifica a relevância de cada palavra de acordo com sua distância às palavras mais frequentes [11]. O comprimento médio dos caminhos mínimos é calculado a partir da distância δ_{ij} , que é o menor número de passos necessários para alcançar o vértice j a partir do vértice i. Dado δ_{ij} , o valor de L para o vértice i é calculado como:

$$L_i = \frac{1}{M-1} \sum_{j \neq i} \delta_{ij} \tag{4}$$

  Em um contexto de análise textual, pode-se concluir que L quantifica a importância da palavra pois mede a distância entre essa palavra e as mais frequentes no texto.

- ***Betweenness***: em redes complexas, quanto maior o número de caminhos pelos quais um vértice ou aresta está presente, maior será a importância desse vértice ou aresta para a rede [40]. Assim, assumindo que todas as interações utilizem os caminhos mínimos entre dois vértices, é possível medir a importância de um vértice ou aresta através da *betweenness centrality*, definida por:

$$B_u = \sum_{ij} \frac{\sigma(i,u,j)}{\sigma(i,j)}, \tag{5}$$

  onde $\sigma(i,u,j)$ é o número de caminhos mínimos entre os vértices i e j que passam pelo vértice ou aresta u, $\sigma(i,j)$ é o número total de caminhos mínimos entre os vértices i e j. Note que a somatória é realizada para todos os pares i e j de vértices. Em uma rede representando texto, palavras muito frequentes tendem a apresentar alto valor de B.

Uma característica textual útil para o reconhecimento de padrões é a distribuição irregular de palavras em diferentes documentos. A análise de palavras recorrentes em documentos específicos pode ser útil para detectar palavras-chaves e até mesmo a autoria de documentos [41, 42]. Esta ideia básica é utilizada na medida *term frequency – inverse document frequency*

(tf-idf) [41]. O termo relativo à frequência é geralmente calculado como:

$$\mathrm{tf}(i) = \frac{N_i}{\sum N_j}, \tag{6}$$

onde N_i representa a frequência da palavra i. Note que $tf(i)$ é definido de forma normalizada para evitar o favorecimento de uma dada palavra em documentos longos. Já o termo *idf* é dado por:

$$\mathrm{idf}(i,t) = \log\left(\frac{\|\mathcal{T}\|}{\|\{\tau \in \mathcal{T} : i \in \tau\}\|}\right) \tag{7}$$

onde $||\mathcal{T}||$ representa o número de documentos na coleção e $\|\{\tau \in \mathcal{T} : i \in \tau\}\|$ quantifica o número de documentos no qual a palavra i ocorre. Porém, em algumas situações, a comparação com um corpus de referência $\mathcal{T}$ não é possível. Isto pode representar um problema para o reconhecimento de autoria quando não há vários exemplos de textos escritos pelos potenciais autores. Problemas deste tipo são tratados fazendo uso do fato de que palavras são distribuídas irregularmente não apenas entre documentos, mas também dentro de um mesmo documento [43].

A quantificação da irregularidade da distribuição de palavras em textos é feita normalmente com medidas estatísticas [44]. Neste projeto, a medida de intermitência [45] será utilizada para quantificar o tempo de recorrência das palavras [43]. Em textos, o tempo de recorrência T_j, para cada palavra i, é definido como o número de palavras entre duas ocorrências sucessivas da palavra i (i.e., ocorrência j e a ocorrência $j+1$) acrescentado de 1. A intermitência da palavra aparece na variância dos T_j's em torno da média. A medida de intermitência é capaz de quantificar a relevância semântica das palavras. Portanto, ela é útil para tarefas de processamento de línguas naturais.

4 Reconhecimento de autoria

Em um típico problema de reconhecimento de autoria, um texto de autoria desconhecida é atribuído a um autor, de um conjunto de vários possíveis autores. As primeiras atividades de reconhecimento de autoria apoiadas por métodos estatísticos tiveram início no século XIX. Uma das primeiras abordagens e que ainda é utilizada em muitos estudos é o uso de modelos probabilísticos [15]. Estes métodos têm por objetivo maximizar a probabilidade *P(x|a)* de um texto x pertencer a um possível autor a. Em [15], os autores analisaram a autoria de centenas de ensaios políticos, conhecidos como *The Federalist Papers*. Neste estudo, os autores introduziram uma grande contribuição para a área ao mostrarem que a frequência de palavras comuns (como *and* e *to* na língua inglesa) produz resultados significativos para a distinção entre os possíveis autores.

Diferentemente dos métodos de reconhecimento de autoria tradicionais, utilizados por linguistas e peritos, o trabalho de [15] iniciou os estudos de atribuição de autoria baseados em estatística. Desde então, muitos trabalhos se dedicaram a definir características que quantificassem o estilo de escrita do autor, conhecido como estilometria [46]. De acordo com

[4], os atributos estilométricos são divididos em algumas categorias, dentre elas destacam-se as características léxicas, baseadas em caracteres, sintáticas e semânticas.

Características léxicas e baseadas em caracteres consideram o texto como uma sequência de palavras ou caracteres, respectivamente. Exemplos de características léxicas são o comprimento das palavras e sentenças, frequência de palavras, a riqueza do vocabulário utilizado (*vocabulary richness*), entre outros. Considerada uma das primeiras tentativas de atribuição de autoria, [47] utilizou-se de medidas de comprimento das sentenças e das palavras para identificar a autoria de textos. Uma importante descoberta com relação ao uso de atributo léxicos é o fato de que palavras comuns (artigos, preposições, pronomes), conhecidas como *function words*, estão entre as melhores características para diferenciar um conjunto de autores [48]. Por serem utilizadas de maneira inconsciente, essas palavras são capazes de capturar o estilo de escrita de cada autor. Além disso, palavras frequentes são responsáveis pelo sucesso da abordagem proposta por físicos em que a distância dos *ranks* de frequência de palavras é usada para computar similaridade entre textos [49]. Outra possível abordagem utilizando características léxicas é extrair as palavras mais frequentes em um texto. Vários trabalhos utilizaram diferentes quantidades de palavras frequentes extraídas do conjunto de dados, desde 100 até 1000 palavras mais frequentes [48, 50].

Para as medidas baseadas em caracteres, um texto é analisado como uma simples sequência de caracteres. Alguns exemplos de medidas são a contagem de caracteres alfabéticos ou numéricos, frequência de letras e sinais de pontuação, entre outros. Uma outra abordagem é extrair a frequência n-gramas em nível de caractere. O sucesso de tais abordagens em quantificar o estilo de escrita sugere ser possível montar uma rede de bigramas de caracteres que são capazes de caracterizar textos. Mais especificamente, no contexto de reconhecimento de autoria, foi observado que a frequência dos caracteres mais comuns representa um bom atributo para discriminar autores, apresentando melhores resultados do que com a utilização de características léxicas [51]. A vantagem da utilização de características baseadas em caracteres é que podem ser utilizadas em basicamente qualquer linguagem natural.

A extração de características sintáticas e semânticas requer uma análise mais elaborada do texto. Informações sintáticas são consideradas mais representativas do que características léxicas, por exemplo. Isso ocorre porque autores costumam utilizar padrões sintáticos inconscientemente. Um indicativo desse poder de representação é o bom desempenho que as palavras comuns (*function words*) apresentam na representação de estilo de escrita, uma vez que essas palavras são encontradas em certas estruturas sintáticas.

Entretanto, as ferramentas atuais de processamento de línguas naturais ainda não são capazes de representar adequadamente o estilo através de informações sintáticas e semânticas, apresentando notável inserção de ruídos. Assim, poucos trabalhos tentam explorar estas características estilométricas, sendo melhor utilizadas como complemento à outras medidas de caracterização, como léxicas ou a nível de caractere.

O reconhecimento de autoria através da utilização de redes de co-ocorrência é abordado em [52]. As medidas extraídas da rede mostraram-se capazes de distinguir os diferentes autores, ressaltando o poder desses valores para a tarefa de reconhecimento de autoria. A partir da combinação de algumas medidas, foi possível agrupar diferentes autores, caracterizando um

estilo comum de escrita.

Uma dependência entre a topologia das redes de co-ocorrência e as características de estilo de autores foi observada no trabalho intitulado *"Comparing intermittency and network measurements of words and their dependence on authorship"* [11]. Neste trabalho, após retirarem do texto as *stopwords*, os autores extraíram medidas de redes complexas e de intermitência com o objetivo de investigar como estas dependem do estilo dos autores. A partir dessas medidas, os autores concluíram que os fatores com maior dependência são a média dos caminhos mínimos e a medida de assimetria da distribuição de intermitência das palavras, conhecida como *skewness*. Com a utilização de alguns métodos de aprendizado de máquina, essas medidas apresentaram uma taxa de acerto de 62.5% para a tarefa de reconhecimento de autoria.

5 Técnicas de reconhecimento de padrão

Diversas técnicas podem ser utilizadas para realizar o reconhecimento de padrões. Entre as mais comuns estão as técnicas *Support Vector Machines*, vizinhos mais próximos e *Naive Bayes* [53]. Antes de descrever esses métodos, considere as definições apresentadas a seguir. O conjunto de treinamento $X_{training} = \{(x_1, y_1), ...(x_l, y_l)\}$ apresenta l tuplas, onde o primeiro componente da i-ésima tupla $x_i = (f_1, ...f_d)$ representa os atributos da instância. O segundo componente, y_i, é o nome da classe da instância. O objetivo de um sistema de classificação supervisionada é aprender o mapeamento $x \mapsto y$ a partir do conjunto de treinamento. Para verificar a acurácia do mapeamento, um conjunto de teste $X_{test} = \{x_{l+1}, ...x_{l+u}\}$ é utilizado.

Algumas técnicas de aprendizado supervisionado são descritas a seguir:

- ***Support Vector Machines***: nesta técnica, os exemplos de treinamento são divididos em diversas regiões do espaço de acordo com suas categorias. Esta divisão é realizada por funções específicas que objetivam maximizar a margem de separação. Desse modo, novas instâncias são então classificadas de acordo com seu mapeamento em uma das regiões de separação.

- **Vizinhos mais próximos**: também conhecida como *k-Nearest Neighbors*. Esta técnica é baseada em um processo de votação realizado sobre as k instâncias mais próximas (k_{nn}) do conjunto de treinamento. Se a maioria das instâncias no conjunto k_{nn} são classificadas com a classe y' então esta classe também é atribuída ao objeto desconhecido.

- ***Naive Bayes***: este método é baseado na regra de decisão que afirma que a correta classe y' de uma instância satisfaz a condição

$$P(y'|f_1, \dots f_d) > P(y_k|f_1, \dots f_d) \tag{8}$$

  para cada $y_k \neq y'$, onde $P(y_k|f_1, \dots f_d)$ é a probabilidade de uma instância ser classificada com a classe y_k dadas as características $F = \{f_1, ...f_d\}$.

Além dos classificadores apresentados acima, outros classificadores também serão utilizados [53]. As versões *fuzzy* desses classificadores também serão consideradas para combinar

técnicas distintas de classificação (ver próxima seção).

6 Atividades propostas e resultados esperados

Nesta seção, as atividades propostas para este projeto são detalhadas, assim como os resultados esperados com essas atividades.

6.1 Extensão da modelagem

Ao longo deste projeto serão propostos modelos alternativos ao modelo de co-ocorrência tradicional. A primeira mudança refere-se à inclusão de *stopwords* (palavras funcionais) na construção da rede pois, tais palavras são úteis em estratégias tradicionais de reconhecimento não baseadas em grafos [4]. Também investigaremos como a conexão de palavras em um maior contexto (não apenas palavras vizinhas como nas redes de adjacência convencionais) afeta a tarefa. Em especial, acreditamos que haverá um tamanho de janela J ótimo onde a classificação será maximizada. Obviamente, o valor de J não deve ser muito alto já que a conexão de palavras mais distantes tornará a modelagem capaz de capturar as características semânticas do texto [22], que a princípio não são úteis nas tarefas de estilometria. Uma vez que a introdução de novos modelos pode mudar a interpretação das medidas, para todos os modelos propostos será realizada uma análise das características das medidas referentes à sua informativade (capacidade de identificação de estruturas em textos) e natureza (i.e. se a medida captura fatores sintáticos ou semânticos).

Um problema comum referente ao tratamento de textos com redes complexas que pode afetar negativamente a capacidade de captura de características estilométricas é a dependência de algumas medidas topológicas com o tamanho do vocabulário (i.e. o número de vértices da rede). Em tarefas onde esta questão é crucial, propomos a normalização das medidas pelo valor esperado em textos aleatórios, fazendo uma analogia direta com o conceito de modularidade em redes, onde o número de inter- ou intra-conexões é normalizado pelo número esperado em redes aleatórias [54]. Para definir esta normalização, considere a seguinte notação. X representa o valor de uma medida (Seção 3.3) obtida em um dado texto. $\mu(X_R)$ representa a média obtida para X em vários textos aleatorizados (i.e., textos em que a ordem das palavras é estabelecida aleatoriamente mantendo-se as distribuições de frequências originais). Desta forma, a medida normalizada $\tilde{X}$ é definida como:

$$\tilde{X} = \frac{X}{\mu(X_R)}. \tag{9}$$

Dado que $\mu(X_R)$ apresenta um erro $\epsilon(\mu(X_R)) = \sigma(X_R)$ devido ao desvio observado nas realizações aleatórias, o erro $\epsilon(\tilde{X})$ observado na variável $\tilde{X}$ é dado por:

$$\epsilon(\tilde{X}) = \left|\frac{d\tilde{X}}{d\mu(X_R)}\right|\epsilon(\mu(X_R)) = \left| - \frac{X}{\mu^2(X_R)}\right|\sigma(X_R) = \frac{\sigma(X_R)}{\mu^2(X_R)}X = \frac{\sigma(X_R)}{\mu(X_R)}\tilde{X}. \tag{10}$$

A normalização efetuada na equação 9 é útil por permitir comparar cada medida como um

modelo nulo. Dessa forma, uma medida fornece informação significativa somente se seu valor $\tilde{X}$ não é próximo de $\tilde{X} = 1$. Para eliminar a complexidade adicional de geração e extração de medidas de textos aleatórios, avaliaremos a possibilidade de se derivar experimentalmente uma expressão analítica que relacione a dependência de $\mu(X_R)$ e $\sigma(X_R)$ com o tamanho do vocabulário para cada medida distinta. Além do uso da equação 9, neste projeto, avaliaremos outras formas de normalização que têm sido utilizadas nas pesquisas de redes complexas [55].

6.2 Uso e concepção de medidas inéditas

Além da proposição dos novos modelos, neste projeto utilizaremos medidas inéditas de redes complexas para análise de fatores estilísticos de textos. Por exemplo, sabe-se que a frequência de palavras específicas representa um fator importante para a caracterização de estilos. Em redes tradicionais de co-ocorrência, a frequência pode ser medida pelo grau dos vértices. Sabendo disto, utilizaremos extensões do conceito de grau que têm sido úteis para aperfeiçoar a caracterização de outros sistemas textuais [56]. A primeira extensão do conceito de grau corresponde ao uso da versão hierárquica desta medida. As medidas hierárquicas são caracterizadas por analisar vizinhanças mais distantes. Portanto, o grau hierárquico de um nó i à distância h, é definido como

$$k_i(h) = N_i(h), \tag{11}$$

onde $N_i(h)$ representa o número de nós que estão a uma distância h do nó i. A Figura 3 ilustra o cálculo do grau hierárquico de um vértice i para diferentes valores de h.

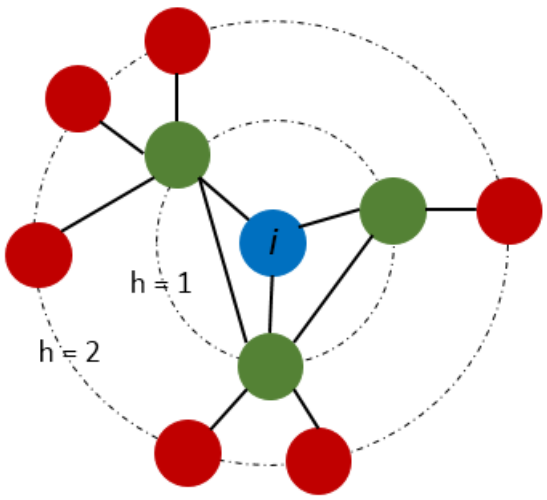


Figura 3: No grafo acima, os três vértices verdes representam aqueles que estão a uma distância h = 1 de i. Portanto, o grau hierárquico $k_i(1) = 3$. Analogamente, os seis vértices vermelhos estão a uma distância h=2 de i, logo $k_i(2) = 6$

Embora a versão hierárquica forneça mais informação topológica local do que simplesmente a medida de grau, esta versão não considera que o acesso a vértices mais distantes

pode ocorrer de maneira irregular. Desta forma, mesmo que um vértice possua grau alto, existe a possibilidade que apenas alguns de seus vários vizinhos sejam efetivamente acessados. Para considerar esta possibilidade, utilizaremos também a medida de acessibilidade [57].

A medida de acessibilidade quantifica o número de nós efetivamente acessíveis a partir de um nó inicial [58]. Para calcular esta medida, considere que $P_h(i,j)$ representa a probabilidade de se alcançar um vértice j a partir de i através de uma caminhada aleatória auto excludente de comprimento h [56]. Neste caso, são considerados os caminhos a partir do vértice i para cada um dos vértices situados no anel concêntrico de distância h. Considerando esta notação, a medida de acessibilidade é dada por

$$a^{(h)}(i) = \exp(-\sum P_h(i,j) \ln P_h(i,j)). \tag{12}$$

Com esta definição, $0 < a^{(h)}(i) \leq N_i(h)$, onde $N_i(h)$ é o número de vértices no anel h. O valor máximo é obtido quando todas as probabilidades de transição para um dado nível concêntrico são iguais. Nesse caso, todos os nós podem ser acessados igualmente. A Figura 4 ilustra as probabilidades de transição e a medida de acessibilidade para duas configurações de redes. Um importante aspecto dessa medida é a propriedade de detectar palavras-chave relevantes do conteúdo abordado [5]. Uma vez que a fração de palavras-chave também se mostrou relevante para a tarefa [11], acredita-se que a introdução desta medida seja capaz de aperfeiçoar a tarefa de reconhecimento de autoria.

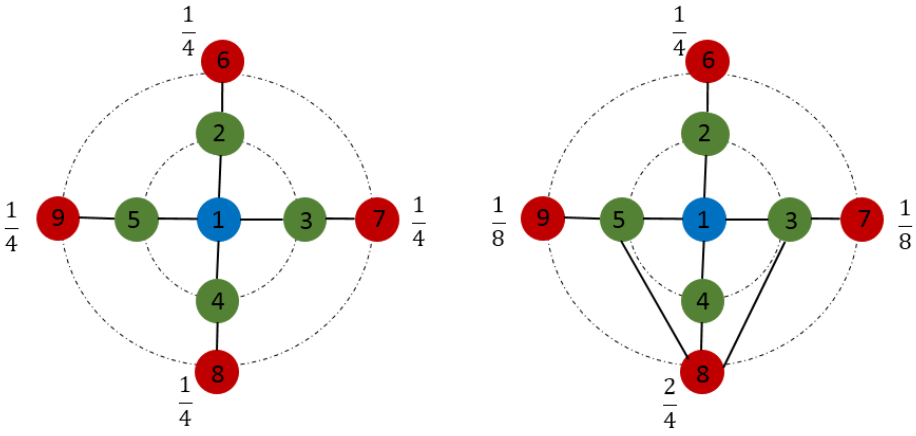


Figura 4: No grafo à esquerda, a probabilidade de transição para cada um de seus vértices do segundo nível concêntrico é a mesma, resultando no maior valor de acessibilidade $a^{(2)}(1) = 4$. No grafo à direita, com a adição de duas novas arestas, as probabilidades de transição se alteram e observa-se uma tendência aparente de acesso ao vértice 8, resultando em um número menor de vértices efetivamente acessados $a^{(2)}(1)$.

Um segundo fator que tem se mostrado importante para a tarefa de reconhecimento de autoria é a heterogeneidade da distribuição de certas quantidades textuais. Por exemplo, em [59], foi mostrado que a heterogeneidade da distribuição de palavras específicas é capaz de identificar estilos de forma significativa. Neste contexto, iremos estudar a heterogeneidade de acesso aos vizinhos de um vértice através de medidas de simetria local récem-propostas.

Em [60], os autores definem duas medidas que quantificam a simetria local de um nó, a simetria *backbone* e simetria *merged*. Estas medidas de simetria são versões normalizadas da acessibilidade [57], onde o número de nós acessíveis é utilizado como fator de normalização. Para calcular essas medidas, as caminhadas aleatórias concêntricas [60] são utilizadas como forma de evitar transições à níveis concêntricos anteriores. Portanto, alterações devem ser feitas na rede de modo que as transições não utilizem arestas dentro de um mesmo nível concêntrico. Na medida de simetria *backbone*, as arestas que conectam nós pertencentes ao mesmo nível concêntrico são eliminadas. Com uma abordagem diferente, a medida *merged* considera essas arestas com custo 0 e os nós conectados por elas são colapsados. A partir dessas alterações na rede, as probabilidades de transição $P_h(i,j)$ são calculadas. A Figura 5 ilustra as alterações realizadas na rede para o cálculo de cada medida e as respectivas probabilidades de transição. Após o cálculo das probabilidades, a simetria *backbone* pode ser calculada da seguinte maneira:

$$Sb_i^{(h)} = \frac{\exp(-\sum P_h(i,j)\ln P_h(i,j))}{|\xi_i^{(h)}|} \tag{13}$$

onde $\xi_i^{(h)}$ representa o conjunto de nós acessíveis que estão a uma distância h do nó i. Para a simetria *merged*, o cálculo a ser realizado é semelhante, porém os pesos das arestas são considerados no cálculo das probabilidades de transição. Essas medidas de simetria conseguem capturar distintos padrões de conectividade [60]. Por enriquecerem a caracterização das redes complexas, acredita-se que essas medidas também sejam capazes de aperfeiçoar a tarefa de reconhecimento de autoria. Na Tabela 2, mostramos a taxa de acerto obtida para o reconhecimento de autoria em um experimento preliminar realizado em uma base de dados contendo 8 autores. As altas taxas de acerto confirmam o potencial da metodologia proposta.

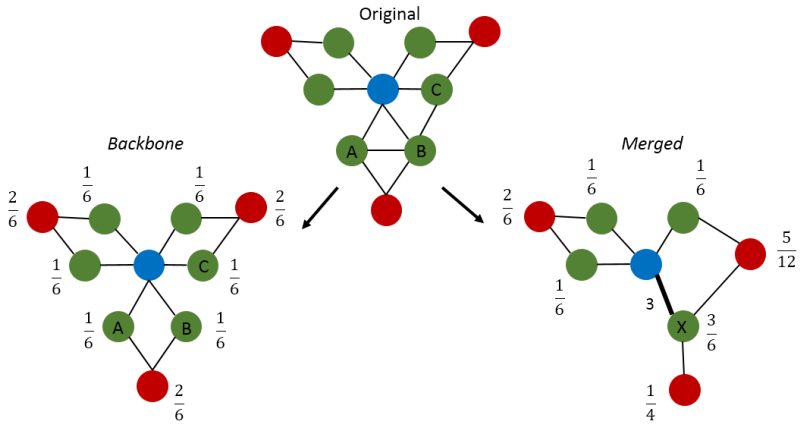


Figura 5: No grafo original acima, as arestas entre A e B e entre B e C conectam vértices pertencentes ao mesmo nível concêntrico. Essas arestas são removidas para realizar o cálculo da simetria *backbone*. Por outro lado, para o cálculo da simetria *merged*, os vértices A, B e C são colapsados em um único vértice X. Pode-se perceber que as diferentes alterações na estrutura do grafo modificam as probabilidades de transição para os vértices no segundo nível concêntrico.

Tabela 2: Taxa de acerto obtida para a tarefa de reconhecimento de autoria usando as medidas de simetria. Os seguintes classificadores foram utilizados: Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), vizinhos mais próximos (kNN) e Naive Bayes (NBY).

Simetria	**SVM**	**MLP**	**KNN**	**NBY**
Merged $h = 2$	75.0%	72.5%	55.0%	42.5%
Merged $h = 3$	70.0%	62.5%	65.0%	40.0%
Merged $h = 4$	82.5%	82.5%	57.5%	42.5%
Backbone $h = 2$	32.5%	32.5%	20.0%	20.0%
Backbone $h = 3$	70.0%	72.5%	57.5%	27.5%
Backbone $h = 4$	70.0%	82.5%	57.5%	42.5%

Uma medida utilizada em [22] e que apresenta bons resultados para a caracterização da topologia de redes é o motivo (*motif*) [61]. Motivos são padrões de conectividade expressos em pequenos blocos de construção (subgrafos). A Figura 6 ilustra alguns exemplos de motivos. O estilo de escrita de cada autor pode gerar diferentes padrões de conectividade na rede textual, por isso essa medida será utilizada neste projeto para auxilar a tarefa de reconhecimento de autoria.

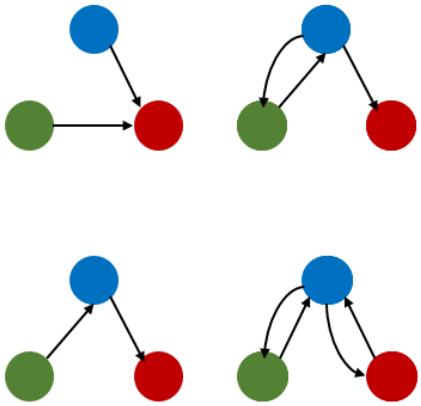


Figura 6: A ilustração acima apresenta 4 tipos diferentes de motivos, mais exemplos podem ser encontrados em [61].

6.3 Combinação com técnicas tradicionais

Os métodos de reconhecimento de autoria baseados em medidas topológicas de redes complexas têm gerado resultados significativos. No entanto, quando estes resultados são comparados com as técnicas tradicionais em Estilometria, geralmente as taxas de acertos obtidas usando apenas topologia são menores do que as obtidas com técnicas tradicionais. Por este motivo, pretendemos combinar as técnicas tradicionais com as baseadas em redes para aperfeiçoar o reconhecimento de estilo em textos. Uma possibilidade de combinação de métodos corresponde ao uso de classificadores híbridos. Neste caso, a classificação pode

ser realizada levando em consideração a probabilidade de uma instância pertencer a uma dada classe considerando cada uma das abordagens. Seja $u_{ij}^{(R)}$ a probabilidade da instância i pertencer à classe j de acordo com a abordagem topológica e $u_{ij}^{(T)}$ a probabilidade da mesma instância pertencer à classe j segundo alguma técnica tradicional. Tais probabilidades podem ser combinadas através de um combinação convexa:

$$u_{ij}^{(H)} = \lambda u_{ij}^{(R)} + (1 - \lambda) u_{ij}^{(T)}, \tag{14}$$

onde $u_{ij}^{(H)}$ representa a probabilidade da instância pertencer à class j de acordo com a combinação de classificadores e λ representa o peso dado à estratégia topológica e está restrito ao intervalo $0 \leq \lambda \leq 1$. Como na equação 14 temos uma probabilidade (p.e. $u_{ij}^{(R)}$) como resultado da classificação, será necessário utilizar classificadores *fuzzy* [62]. A princípio, o melhor valor a ser escolhido para λ não está claro. Por este motivo, também estudaremos formas de prever a melhor configuração para o parâmetro λ. Uma abordagem similar à proposta na equação 14 foi implementada com sucesso em [63] para desambiguar nomes de autores em redes de colaboração. Tal abordagem deve ser utilizada como base para este projeto.

Além da combinação fornecida na equação 14, existe também a possibilidade de se classificar inicialmente com a abordagem tradicional. Em seguida, a abordagem topológica poderia ser utilizada para classificar as instâncias cujas classes inferidas pelo método tradicional apresentaram alto grau de incerteza. Formas complementares de combinação de classificadores supervisionados também serão investigadas neste projeto.

A seguir, listamos alguns exemplos de atributos que podem ser utilizados para compor a classificação baseada em técnicas estatísticas tradicionais.

- **Palavras funcionais**: a frequência de palavras funcionais específicas representa uma das principais marcas estilísticas utilizadas para realizar o reconhecimento de autoria. Portanto, a frequência de tais palavras pode ser utilizada como sendo um atributo relevante para a abordagem topológica.

- **Caracteres**: a frequência de caracteres (ou bigramas) específicos também representa um fator relevante para a classificação estilométrica de textos. A consideração deste tipo de atributo nos classificadores tradicionais também é direta, já que a frequência de cada caractere (ou bigrama) pode ser utilizada como atributo. A vantagem do uso deste tipo de atributo é que ele não necessita de nenhuma informação textual profunda. Além disso, como mencionado na Seção 4, existe a possibilidade de criar redes de bigramas de caracteres.

- **Estrutura textual**: em vários estudos linguísticos comprova-se a existência de certos padrões organizacionais de textos. Sabe-se, por exemplo, que correlações de longo alcance ocorrem devido ao mapeamento de ideias de forma linear [64]. Outra descoberta relevante refere-se à entropia da distribuição de palavras, que comprova que certas palavras possuem distribuições espaciais de acordo com a função exercida [22]. Uma vez que ambos os fatores já foram utilizados em tarefas de identificação de língua,

adaptaremos estes métodos para capturar as variações estilísticas mais sutis que podem diferenciar autores.

6.4 Aplicações adicionais

Além das atividades apresentadas anteriormente, a concepção de novos métodos, medidas e combinações de classificadores podem também auxiliar o desenvolvimento de tarefas correlatas à tarefa de reconhecimento de autoria. A seguir, listamos algumas dessas tarefas:

- **Inconsistência estilística:** quando duas ou mais pessoas escrevem um documento, sendo cada um responsável por uma parte do texto, podem surgir inconsistências estilísticas devido à múltipla autoria, o que pode gerar textos com baixa coerência. A identificação de inconsistência estilística também pode ser utilizada para detectar fraudes em documentos que deveriam ser escritos por um único autor. Este problema pode ser identificado, por exemplo, dividindo-se o texto em pedaços e transformado cada pedaço em uma rede. Este tipo de divisão mostrou-se viável no estudo realizado em [39].

- **Detecção de plágio:** as técnicas de detecção de plágio tradicionais baseiam-se principalmente no conteúdo semântico dos textos, de uma maneira similar às medidas de avaliação de qualidade de traduções automáticas [65]. As técnicas baseadas em análise de topologia de redes complexas poderiam ser úteis, por exemplo, para identificar tentativas de mascaramento de plágios. Se uma ou mais palavras da fonte original são substituídas por sinônimos, o *overlapping* com a fonte diminui. Este tipo de fraude não afetaria nossa técnica pois uma mudança global de rótulos de palavras específicas não afetaria a organização dos textos.

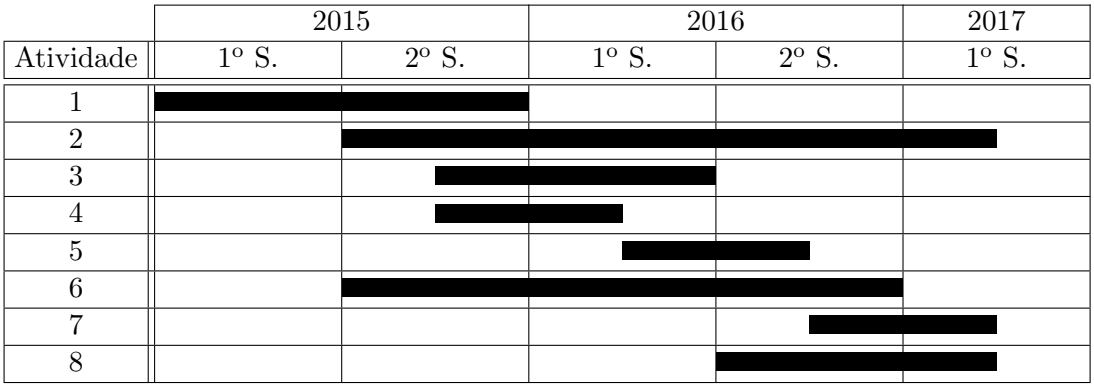
- **Identificação de fraudes em reconhecimento**: estudos recentes mostram que os métodos de detecção de autoria tradicionais não são robustos à ataques de autores que tentam mascarar estilos [17]. Neste sentido, pretendemos analisar a robustez das redes complexas com relação a estes ataques.

7 Plano de Atividades e Cronograma Previsto

As seguintes atividades estão previstas para os 24 meses de execução deste projeto, com início em fevereiro de 2015 (data da matrícula da candidata como aluna regular de mestrado do Programa CCMC do ICMC/USP) e defesa prevista para fevereiro de 2017. O cronograma prevê a obtenção dos créditos de aula obrigatórios do programa, a realização do projeto de pesquisa e os exames de qualificação e defesa.

1. **Disciplinas da pós-graduação:** No primeiro semestre de 2015, a aluna pretende cursar 31 dos 51 créditos de disciplinas requeridos para obtenção do título de mestre em Ciências de Computação e Matemática Computacional, representando cinco disciplinas do programa;

Tabela 3: Cronograma de Atividades.



2. **Revisão bibliográfica:** Revisão do estado da arte em processamento de línguas naturais e reconhecimento de autoria, incluindo os estudos devotados ao tratamento de textos a partir de conceitos de redes complexas;

3. **Aplicabilidade e Obtenção de dados:** Levantamento de possíveis aplicações que podem ser melhoradas com a utilização dos modelos alternativos a serem propostos. Além do estudo de técnicas e ferramentas tradicionais em reconhecimento de autoria;

4. **Qualificação:** Preparação e apresentação do exame de qualificação, conforme determina as normas do Programa de Pós-graduação;

5. **Estágio de Pesquisa no Exterior:** A aluna pretende realizar um estágio de pesquisa no exterior com duração máxima de 6 meses. Foi realizado o primeiro contato com a pesquisadora no exterior, Kathleen McKeown, professora e diretora do *Institute for Data Sciences and Engineering* da Universidade de Columbia, em Nova York - EUA, e aguarda-se o aceite da pesquisadora;

6. **Desenvolvimento e Implementação:** Desenvolvimento e implementação da proposta do projeto;

7. **Avaliação de Resultados:** Avaliação dos resultados e comparação com ferramentas e técnicas existentes;

8. **Trabalhos Científicos:** Redação de artigos científicos, relatórios, participação de congressos e escrita da tese final de mestrado bem como sua apresentação para uma banca avaliadora.

Referências

[1] R. Albert and A.-l. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys*, 2002.

[2] R. Mihalcea and D. Radev, *Graph-based natural language processing and information retrieval*. Cambridge; New York: Cambridge University Press, 2011.

[3] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[4] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, pp. 538–556, Mar. 2009.

[5] D. R. Amancio, M. G. V. Nunes, O. N. Oliveira and L. da F. Costa, "Extractive summarization using complex networks and syntactic dependency," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 4, pp. 1855 – 1864, 2012.

[6] H. Liu, "Statistical properties of chinese semantic networks," *Chinese Science Bulletin*, vol. 54, no. 16, pp. 2781–2785, 2009.

[7] G. Ludueña, M. Behzad, and C. Gros, "Exploration in free word association networks: models and experiment," *Cognitive Processing*, vol. 15, no. 2, pp. 195–200, 2014.

[8] R. F. i Cancho and R. V. Solé, "The small world of human language," *Proceedings of The Royal Society of London. Series B, Biological Sciences*, vol. 268, pp. 2261–2266, 2001.

[9] R. F. i Cancho, R. V. Solé, and R. Köhler, "Patterns in syntactic dependency networks," *Phys. Rev. E*, vol. 69, p. 051915, May 2004.

[10] D. R. Amancio, O. N. Oliveira, and L. d. F. Costa, "Identification of literary movements using complex networks to represent texts," *New Journal of Physics*, vol. 14, no. 4, p. 043029, 2012.

[11] D. R. Amancio, E. G. Altmann, O. N. Oliveira, and L. d. F. Costa, "Comparing intermittency and network measurements of words and their dependence on authorship," *New Journal of Physics*, vol. 13, p. 123024, Dec. 2011.

[12] T. C. Silva and D. R. Amancio, "Word sense disambiguation via high order of learning in complex networks," *EPL (Europhysics Letters)*, vol. 98, no. 5, p. 58001, 2012.

[13] T. Grant, "Quantifying evidence in forensic authorship analysis," *International Journal of Speech Language and the Law*, vol. 14, no. 1, 2007.

[14] A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group web forum messages," *IEEE Intelligent Systems*, vol. 20, pp. 67–75, 2005.

[15] F. Mosteller and D. L. Wallace, *Inference and Disputed Authorship: The Federalist Papers*. Reading, Mass.: Addison-Wesley, 1964.

[16] P. Juola, "Authorship attribution," *Found. Trends Inf. Retr.*, vol. 1, pp. 233–334, Dec. 2006.

[17] M. R. Brennan and R. Greenstadt, "Practical attacks against authorship recognition techniques.," in *IAAI* (K. Z. Haigh and N. Rychtyckyj, eds.), AAAI, 2009.

[18] R. A. J. Matthews and T. V. N. Merriam, "Neural computation in stylometry i: An application to the works of shakespeare and fletcher," *Literary and Linguistic Computing*, vol. 8, no. 4, pp. 203–209, 1993.

[19] F. J. Tweedie, S. Singh, and D. I. Holmes, "Neural network applications in stylometry: The federalist papers.," *Computers and the Humanities*, vol. 30, no. 1, pp. 1–10, 1996.

[20] G. Frantzeskou, E. Stamatatos, S. Gritzalis, and S. K. Katsikas, "Effective identification of source code authors using byte-level information.," in *ICSE* (L. J. Osterweil, H. D. Rombach, and M. L. Soffa, eds.), pp. 893–896, ACM, 2006.

[21] C. E. Chaski, "Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations," *International Journal of Digital Evidence*, vol. 4, 2005.

[22] D. R. Amancio, E. G. Altmann, D. Rybski, O. N. Oliveira, and L. d. F. Costa, "Probing the statistical properties of unknown texts: application to the Voynich Manuscript.," *PloS one*, vol. 8, p. e67310, Jan. 2013.

[23] J. L. Gross and J. Yellen, *Graph Theory and Its Applications, Second Edition (Discrete Mathematics and Its Applications)*. Chapman & Hall/CRC, 2005.

[24] B. Bollobás, *Graph Theory*. North-Holland Mathematics Studies, Elsevier Science, 1982.

[25] B. Bollobas, *Modern Graph Theory*. Graduate Texts in Mathematics, Springer New York, 1998.

[26] B. Bollobás, *Random graphs*. Cambridge University Press, 2001.

[27] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, pp. 661–703, Nov. 2009.

[28] D. J. Watts and S. H. Strogatz, "Collective dynamics of /'small-world/' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[29] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[30] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, p. 208701, Oct 2002.

[31] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, pp. 3200–3203, Apr 2001.

[32] C. Castellano, S. Fortunato, and V. Loreto, "Statistical physics of social dynamics," *Rev. Mod. Phys.*, vol. 81, pp. 591–646, May 2009.

[33] D. R. Amancio, S. M. Aluísio, O. N. O. Jr., and L. da Fontoura Costa, "Complex networks analysis of language complexity," *CoRR*, vol. abs/1302.4490, 2013.

[34] I. Grabska-Gradzinska, A. Kulig, J. Kwapien, and S. Drozdz, "Complex Network Analysis of Litarary and Scientific exts," *International Journal of Modern Physics C*, vol. 23, p. 1250051, July 2012.

[35] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 1st ed., 2009.

[36] R. M. Roxas and G. Tapang, "Prose and Poetry Classification and Boundary Detection Using Word Adjacency Network Analysis," *International Journal of Modern Physics C*, vol. 21, pp. 503–512, 2010.

[37] E. Alvarez-Lacalle, B. Dorow, J. P. Eckmann, and E. Moses, "Hierarchical structures induce long-range dynamical correlations in written texts," *Proc Natl Acad Sci U S A*, vol. 103, pp. 7956–7961, May 2006.

[38] D. R. Amancio, O. N. O. Jr., and L. da F. Costa, "Structure-semantics interplay in complex networks and its effects on the predictability of similarity in texts," *CoRR*, vol. abs/1303.0350, 2013.

[39] D. R. Amancio, "Probing the topological properties of complex networks modeling short written texts," *CoRR*, vol. abs/1412.8504, 2014.

[40] L. daF. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: A survey of measurements," *Advances in Physics*, vol. 56, pp. 167–242, January 2007.

[41] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[42] F. Mosteller and D. L. Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, 1984.

[43] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, "Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words," *PLoS ONE*, vol. 4, p. e7678, 11 2009.

[44] M. A. Montemurro and D. H. Zanette, "Keywords and co-occurrence patterns in the voynich manuscript: An information-theoretic analysis," *PLoS ONE*, vol. 8, p. e66344, 06 2013.

[45] J. P. Herrera and P. A. Pury, "Statistical keyword detection in literary corpora," *CoRR*, vol. abs/cs/0701028, 2007.

[46] D. I. Holmes, "Authorship attribution," *Computers and the Humanities*, vol. 28, no. 2, pp. 87–106, 1994.

[47] T. C. Mendenhall, "The characteristic curves of composition," *Science*, vol. ns-9, no. 214S, pp. 237–246, 1887.

[48] J. F. Burrows, "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style," *Literary and Linguistic Computing*, vol. 2, pp. 61–70, 1987.

[49] R. Ferrer i Cancho, "Euclidean distance between syntactically linked words," *Phys. Rev. E*, vol. 70, p. 056135, Nov 2004.

[50] E. Stamatatos, "Authorship attribution based on feature set subspacing ensembles.," *International Journal on Artificial Intelligence Tools*, no. 5, pp. 823–838, 2006.

[51] M. Jankowska, E. Milios, and V. Keselj, "Author verification using common n-gram profiles of text documents," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin City University and Association for Computational Linguistics, 2014.

[52] L. Antiqueira, T. A. S. Pardo, M. G. V. Nunes, O. N. Oliveira Jr., and L. F. Costa, "Some issues on complex networks for author characterization," in *Fourth Workshop in Information and Human Language Technology (TIL'06) in the Proceedings of International Joint Conference IBERAMIA-SBIA-SBRN* (S. O. Rezende and A. C. R. da Silva Filho, eds.), (Ribeirão Preto, Brazil), ICMC-USP, October 23-28 2006.

[53] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2000.

[54] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75 – 174, 2010.

[55] B. C. M. van Wijk, C. J. Stam, and A. Daffertshofer, "Comparing brain networks of different size and connectivity density using graph theory," *PLoS ONE*, vol. 5, p. e13701, 10 2010.

[56] B. A. N. Travençolo, M. P. Viana, and L. da Fontoura Costa, "Border detection in complex networks," *New Journal of Physics*, vol. 11, no. 6, p. 063019, 2009.

[57] B. Travençolo and L. da F. Costa, "Accessibility in complex networks," *Physics Letters A*, vol. 373, no. 1, pp. 89 – 95, 2008.

[58] M. P. Viana, J. a. L. B. Batista, and L. d. F. Costa, "Effective number of accessed nodes in complex networks," *Phys. Rev. E*, vol. 85, p. 036105, Mar 2012.

[59] D. R. Amancio, "Authorship recognition via fluctuation analysis of network topology and word intermittency," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2015, no. 3, p. P03005, 2015.

[60] F. N. Silva, C. H. Comin, T. K. D. M. Peron, F. A. Rodrigues, C. Ye, R. C. Wilson, E. Hancock, and L. F. Costa, "Concentric Network Symmetry," pp. 1–15, 2014.

[61] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[62] L. I. Kuncheva, *Fuzzy Classifier Design*, vol. 49 of *Studies in Fuzziness and Soft Computing*. Springer, 2000.

[63] D. Amancio, O. Oliveira jr, and L. da F. Costa, "Topological-collaborative approach for disambiguating authors' names in collaborative networks," *Scientometrics*, vol. 102, no. 1, pp. 465–485, 2015.

[64] E. G. Altmann, G. Cristadoro, and M. D. Esposti, "On the origin of long-range correlations in texts," *Proceedings of the National Academy of Sciences*, vol. 109, no. 29, pp. 11582–11587, 2012.

[65] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, (Stroudsburg, PA, USA), pp. 311–318, Association for Computational Linguistics, 2002.