UTILIZAÇÃO DO ALGORITMO DE AGRUPAMENTO DE MARKOV (MCL) PARA A IDENTIFICAÇÃO DE CONLUIOS NO MERCADO DE CAPITAIS

Jorge Ivan Hmeljevski[1], Alexandre Leopoldo Gonçalves[2], José Leomar Todesco[3]

Abstract. *This article demonstrates the use of the Markov clustering algorithm (MCL) for the identification of collusions in the capital market. The demonstration covers the description of the algorithm and a case study illustrating its application. The case study was carried out in the context of the supervision that the federal regulator of the Brazilian capital market exercises over the operations on the country's stock exchange. The obtained result proves that the application of this algorithm has the potential to improve the quality of the data analysis performed by the supervisory bodies in order to identify investors who, acting in collusion, commit irregularities in the capital market.*

Keywords: *mcl; clustering; data mining; supervision; stock exchange.*

Resumo. *Este artigo demonstra a utilização do algoritmo de agrupamento de Markov (MCL) para a identificação de conluios no mercado de capitais. A demonstração abrange a descrição do algoritmo e um estudo de caso exemplificando a sua aplicação. O estudo de caso foi realizado no contexto da supervisão que o regulador federal do mercado de capitais brasileiro exerce sobre as operações na bolsa de valores do país. O resultado obtido comprova que a aplicação deste algoritmo tem o potencial para aprimorar a qualidade das análises de dados realizadas pelos órgãos supervisores visando identificar investidores que, atuando em conluio, cometem irregularidades no mercado de capitais.*

Palavras-Chave: *mcl; agrupamento; mineração de dados; supervisão; bolsa de valores.*

[1] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento– Universidade Federal de Santa Catarina (UFSC) Florianópolis – SC– Brazil. Email: jorgeih@gmail.com
[2] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento– Universidade Federal de Santa Catarina (UFSC) Florianópolis – SC– Brazil. Email: a.l.goncalves@ufsc.br
[3] Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento– Universidade Federal de Santa Catarina (UFSC) Florianópolis – SC– Brazil. Email: jose.todesco@ufsc.br

1 INTRODUÇÃO

As irregularidades no mercado de capitais frequentemente são realizadas em conluio, ou seja, através da combinação entre duas ou mais partes atuantes no mercado, sejam estas partes formadas por investidores ou intermediários das operações. Nestes casos, indivíduos se reúnem em grupos com o objetivo de voluntariamente influenciar os negócios na bolsa de valores visando obter vantagens para si, invariavelmente em detrimento de prejuízos irregularmente impostos a outros investidores.

Tais práticas irregulares, quando perpetuadas no longo prazo, trazem prejuízos não só para os investidores diretamente vitimados pelas operações, mas também para o mercado de capitais, para a economia, para o governo e para a sociedade de modo geral. Em decorrência, as iniciativas para a identificação e a punição dos envolvidos em tais irregularidades são de importância central nas atividades dos reguladores e supervisores do mercado de capitais.

Para ilustrar o problema, considera-se o exemplo de um investidor, o qual já possui em custódia uma grande quantidade de determinada ação. Com o objetivo de vender essas ações a um preço vantajoso, este investidor poderia realizar diversas compras de lotes mínimos (lotes padrão de 100 ações) visando, através de poucas ações, gerar uma falsa percepção de liquidez e de demanda por um ativo normalmente pouco negociado no mercado. O vendedor, contraparte destes negócios, poderia ser outro investidor previamente combinado com o comprador (conluio) ou até mesmo ser o próprio comprador, mas vendendo ações através de outro veículo de investimento, como uma empresa limitada sobre a qual ele tenha controle. O aumento artificial do número de negócios e dos preços praticados passa a atrair mais investidores interessados em negociar, de boa fé, com aquele ativo. Neste exemplo, o novo fluxo de investidores tende a pressionar os preços a subir. Em determinado momento, o investidor que iniciou o processo de manipulação realiza a venda de uma grande quantidade de ações, a um preço relativamente alto. Após conseguir seu objetivo e sair do mercado, o preço do ativo, agora sem a demanda compradora artificial, tende a cair novamente para um equilíbrio de mercado num patamar inferior, implicando em prejuízo para os demais investidores atraídos para os negócios com aquela ação.

Obviamente se trata de um exemplo hipotético, mas que ilustra, dentre muitas outras possibilidades, como o conluio entre compradores e vendedores poderia ser utilizado numa tentativa de manipular os preços de ações.

Os dados das operações nos pregões da bolsa de valores normalmente estão disponíveis para serem analisados por parte da supervisão realizada pelos reguladores e fiscalizadores do

mercado de capitais. No entanto, o grande volume de dados envolvido nas operações da bolsa implica que negócios fraudulentos possam passar despercebidos numa análise meramente visual desses dados, mesmo porque, os envolvidos nesses negócios intencionalmente buscam formas de ocultar suas operações da fiscalização das corretoras de valores, da bolsa e do regulador de mercado.

Portanto, torna-se imperativo buscar formas mais eficazes de detectar e combater essas irregularidades, ampliando a aplicação das análises estatísticas e da mineração de dados (*data mining*) por parte da supervisão realizada pelos reguladores e fiscalizadores da bolsa de valores.

Este trabalho apresenta a aplicação de um algoritmo de agrupamento dos dados da bolsa de valores visando identificar conluios entre investidores. O algoritmo de agrupamento utilizado é o de Markov (MCL), conforme proposto por Van Dongen (2000a). O estudo foi realizado no contexto das atividades de supervisão realizadas pela Comissão de Valores Mobiliários (CVM) sobre as operações realizadas na Brasil Bolsa Balcão (B3).

Além desta introdução, na segunda seção deste trabalho são apresentadas mais informações a respeito da CVM, ou seja, do contexto no qual se desenvolve o estudo de caso. Na terceira seção são apresentados os procedimentos metodológicos aplicados, inclusive maiores detalhes sobre o algoritmo MCL e sobre os dados utilizados. A quarta seção discute os resultados obtidos e na quinta seção se conclui o artigo.

2 COMISSÃO DE VALORES MOBILIÁRIOS

O desenvolvimento deste trabalho se deu através de um estudo de caso na CVM. A CVM é uma autarquia em regime especial, vinculada ao Ministério da Fazenda brasileiro e criada com o princípio básico da defesa do mercado de valores mobiliários em geral (mercado de capitais), mas particularmente dos interesses dos investidores, em especial dos acionistas minoritários (Brasil, 1976).

Alinhada ao seu mandato legal, a missão da CVM é desenvolver, regular e fiscalizar o mercado de valores mobiliários como instrumento de captação de recursos para as empresas, protegendo o interesse dos investidores e assegurando ampla divulgação das informações sobre os emissores e seus valores mobiliários (Brasil, 2013).

A defesa dos interesses dos acionistas minoritários implica em garantir a regularidade dos negócios realizados no mercado de capitais. Esta garantia decorre, num primeiro momento, da regulação imposta pela CVM aos integrantes desse mercado, mas num segundo momento é consequência da função fiscalizadora da CVM para fazer valer sua regulação, inclusive através

da aplicação de penalidades administrativas àqueles que descumprem as regras por ela estabelecidas.

A estrutura organizacional da CVM contempla diversas superintendências especializadas em suas respectivas funções, dentre as quais a Superintendência de Relações com o Mercado e Intermediários (SMI). As responsabilidades da SMI, por sua vez, são operacionalizadas por suas gerências. A Gerência de Acompanhamento de Mercado (GMA), por exemplo, é a gerência da SMI responsável pela supervisão das operações na bolsa de valores.

O trabalho realizado pela GMA consiste basicamente de análises intensivas em conhecimento. Por exemplo, diante da detecção de uma possível irregularidade nas operações realizadas na bolsa, inicia-se um processo de análise de dados de operações, de investidores e de informações de companhias. Estas análises visam descartar a hipótese de irregularidade ou comprová-la, caso em que os investidores ou intermediários envolvidos serão acusados e terão que responder pelas irregularidades cometidas.

Assim, na execução de suas atribuições profissionais, os analistas da GMA frequentemente analisam os dados das operações realizadas na B3 visando verificar a existência de irregularidades nesses negócios. Em geral, estas análises se dão sobre dados estruturados e utilizam técnicas estatísticas e de extração de conhecimento de bancos de dados.

A qualidade das análises realizadas implica em sérias consequências quanto à adequada responsabilização administrativa ou até mesmo penal dos investidores que cometem irregularidades nas operações realizadas na bolsa. A efetividade deste processo impacta a segurança do mercado de capitais de modo geral, pois está diretamente relacionada com a credibilidade da CVM em relação à sua capacidade de desempenhar seu papel fiscalizador. Em decorrência, se justifica o esforço em melhorias da qualidade das análises dos dados através da aplicação de algoritmos de mineração de dados, como é o caso do MCL.

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia utilizada fundamenta-se na representação dos negócios realizados entre os investidores na bolsa na forma de um grafo (o grafo de fluxo de vendas). Para chegar a esta representação, foi necessário realizar um pré-processamento dos dados de operações que tipicamente não se encontram nesse formato. A partir da adequada organização dos dados, foi possível aplicar o algoritmo MCL para realizar agrupamentos relevantes entre os investidores. O estudo de caso considerou dados reais de negócios na bolsa de valores, dentre os quais foram

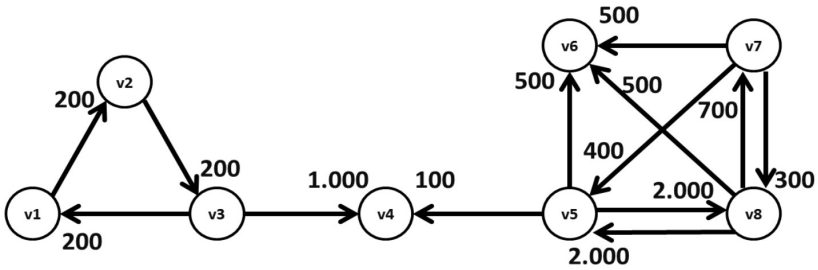
injetados dados fictícios candidatos a irregularidades. Estes aspectos dos procedimentos metodológicos são detalhados nas próximas subseções deste trabalho.

3.1 GRAFO DO FLUXO DE VENDAS

Palshikar e Apte (2005) e, posteriormente, Islam et al. (2009), apresentaram trabalhos nos quais as operações realizadas na bolsa de valores são tratadas como um grafo direcional, o qual foi chamado por estes autores de grafo do fluxo de ações. Conforme proposto por estes autores, o grafo de fluxo de ações G para uma determinada ação S pode ser denotado por GS = (V, A, ϕ), onde V é o conjunto de vértices de G (cada vértice representado pelo identificador único de um investidor – seu *id*), A é o conjunto de arestas direcionais de G (as quais representam as quantidades de ações vendidas pelos investidores) e ϕ é a função que associa para cada aresta um vértice fonte (vendedor de ações) e um vértice alvo (comprador de ações).

O grafo de fluxo de ações pode apresentar arestas paralelas, pois as vendas de ações podem ocorrer, por exemplo, de um investidor A para um investidor B, mas também no sentido oposto. O grafo não apresenta laços (*self-loops*), ou seja, um investidor não vende ações para si mesmo. Além disso, o grafo não apresenta vértices (investidores) isolados, pois um investidor sempre estará associado a, no mínimo, uma aresta estabelecendo uma ligação com outro investidor (uma venda). Por fim, vale destacar que o grafo representa o somatório da quantidade de ações vendidas num determinado período de análise. A Figura 1 apresenta um exemplo desse tipo de grafo.

Figura 1 – Grafo do fluxo de ações.



Fonte: os autores (2017).

Na Figura 1 o investidor v3, no período analisado, vendeu um total de 200 ações para o investidor v2 e um total de 1.000 ações para o investidor v4. Por outro lado, o investidor v1 vendeu um total de 200 ações para o investidor v3 e assim por diante.

Embora o modelo proposto por Palshikar e Apte (2005) e por Islam et al. (2009) represente adequadamente o fluxo da **quantidade de ações** negociadas entre pares de

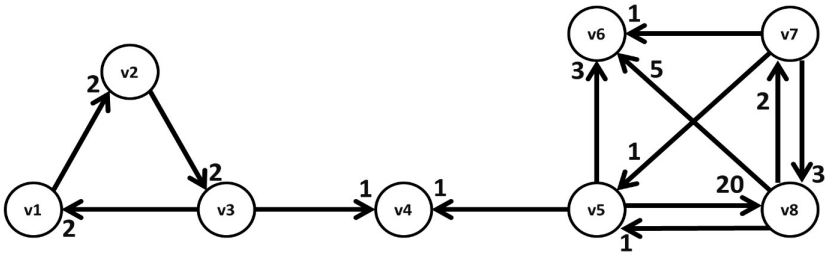
investidores na bolsa de valores, em muitos casos, ao invés do número total de ações negociadas entre dois investidores, **o número total de negócios** (vendas) realizados entre eles pode ser um melhor indicador da existência de um conluio nas operações.

Considere-se, como exemplo simplificado, os negócios realizados entre os investidores v1, v2 e v3 apresentados na Figura 1. Estes negócios poderiam representar uma dinâmica fraudulenta, na qual, conforme previamente acordado entre eles, o investidor v3 vende 100 ações para o v2, em seguida o v2 vende as mesmas 100 ações para o investidor v1 e este as vende novamente para o investidor v3. Este ciclo se repete mais uma vez e assim fica caracterizada a venda de 200 ações representada no grafo. A aparente elevação na quantidade de negócios e a valorização do preço do ativo passam a ser atraentes para outros investidores. Assim, em determinado momento o investidor v3 consegue vender, num só negócio e com um preço mais elevado, 1.000 ações para o investidor v4, o qual realizou sua compra de boa-fé.

No exemplo apresentado, a irregularidade reside nos negócios realizados entre v1, v2 e v3 (a circulação fraudulenta de 100 ações). No entanto, ao estabelecer que as arestas representam o **número de ações** vendidas de um investidor para outro, inadequadamente atribui-se uma relação mais significativa entre o negócio dos investidores v1 e v4 (apenas uma venda de 1.000 ações) do que entre os negócios realizados pelos investidores v1, v2 e v3. Por outro lado, considerando que as arestas representem a **quantidade de negócios** (vendas) entre as partes, a ocorrência de duas vendas entre v1, v2 e v3 (100 ações em cada venda) e de apenas uma venda de v1 para v4 (de 1.000 ações) permite atribuir maior importância para os negócios entre v1, v2 e v3.

Assim, neste trabalho se opta por uma abordagem diferente da apresentada por Palshikar e Apte (2005) e por Islam et al. (2009). Enquanto estes autores modelaram a relação entre os investidores considerando o número de ações vendidas de um investidor para outro, aqui se utiliza o número de negócios (vendas) realizados num determinado período. Com isso, o grafo apresentado na Figura 1, passa a ser representado conforme a Figura 2, ou seja, considerando que cada aresta representa a quantidade de vendas realizadas entre as partes.

Figura 2 – Grafo do fluxo de vendas.



Fonte: os autores (2017).

3.2 APLICAÇÃO DO ALGORÍTMO MCL

As cadeias de Markov são um caso particular de processo estocástico com estados discretos utilizado na computação com diversas finalidades, mas foi Van Dongen (2000a) o propositor de um algoritmo de agrupamento, o MCL, baseado em suas propriedades.

No grafo da Figura 2 é possível constatar que alguns vértices apresentam mais ligações diretas entre si do que com outros vértices. Por exemplo, os vértices v5, v6, v7 e v8 apresentam várias ligações diretas entre si, mas nenhuma ligação direta com os vértices v1, v2 e v3 (apenas ligações indiretas). Considerando, por exemplo, que todas as probabilidades de transição do grafo apresentado na Figura 2 fossem iguais, um passeio aleatório a partir do vértice v8 apresentaria maior probabilidade de permanecer entre os vértices v5, v6, v7 e v8 e menor probabilidade de chegar até os vértices v1, v2 e v3.

Assim, o MCL se baseia na realização de passeios aleatórios pelo grafo, os quais são feitos considerando as probabilidades de transição entre os vértices e a partir dos quais é possível descobrir em que conjuntos de vértices (agrupamento) o fluxo que percorre o grafo tende a se concentrar.

Para a utilização do MCL, inicialmente é necessário representar o grafo na forma de uma matriz de adjacência. Em seguida, a matriz de adjacência é somada com sua respectiva matriz identidade. Este procedimento corresponde à adição de um laço unitário de cada vértice para consigo mesmo (*self-loop*). Embora opcional, a adição deste *self-loop* é indicada para otimizar o funcionamento do algoritmo (Van Dongen, 2000a).

A partir da matriz de adjacência acrescida de valores unitários em sua diagonal principal, o algoritmo MCL cria uma matriz de Markov, ou seja, é realizada uma normalização da matriz de adjacência. Esta normalização consiste na soma de todos os elementos de cada coluna e na divisão de cada elemento da coluna pela soma obtida. Com isto, os valores de cada coluna da matriz podem ser interpretados como a probabilidade de que um passeio aleatório ocorra de um vértice de origem para um vértice alvo. Por exemplo, a partir do investidor v3 existem três possibilidades de transição no grafo apresentado na Figura 2: de v3 para v1, de v3 para v2 ou de v3 para v4. A normalização dos valores das arestas que representam as possíveis transições resulta que as probabilidades de, num passeio aleatório, partir do vértice v3 e chegar aos demais vértices são dadas respectivamente por $P(Xn+1 = v1 \mid Xn = v3) = 2/5$; $P(Xn+1 = v2 \mid Xn = v3) = 2/5$; e $P(Xn+1 = v4 \mid Xn = v3) = 1/5$.

Em seguida, o algoritmo passa a alternar, respectivamente, duas operações chamadas de **expansão** e de **inflação** e iterativamente recalcular as probabilidades de transição entre os vértices do grafo.

A operação de **expansão** consiste numa potenciação da matriz de transição. A potência de expansão *e* é um dos parâmetros do algoritmo, mas como padrão o algoritmo utiliza *e* = 2 (eleva-se a matriz ao quadrado). O operador de expansão é responsável por permitir que o fluxo do caminho aleatório chegue a diferentes regiões do grafo (Van Dongen, 2000a).

A aplicação do operador de **inflação** começa por um produto de Hadamard-Schur, ou seja, pela elevação individual dos elementos da matriz estocástica a um expoente positivo *h*: $(aij)^h$. A operação de inflação tem como efeito o fortalecimento das ligações entre os vértices mais ligados entre si e o enfraquecimento dos vértices menos ligados entre si (Van Dongen, 2000a). Após a operação de inflação, a matriz resultante é normalizada novamente.

A sucessiva realização das operações de normalização, expansão e inflação ao longo de um número suficiente de iterações leva o algoritmo à convergência, quando a matriz de transições assume a forma de uma matriz idempotente, ou seja, uma matriz que multiplicada por si mesma resulta na própria matriz (M x M = M). O algoritmo pode ser resumidamente descrito conforme o Quadro 1.

Quadro 1 – Algoritmo MCL

1. Entradas: matriz de adjacência M e os expoentes de expansão (*e*) e de inflação (*h*);
2. Fazer M1 = M + IM // Adiciona laços a M somando-a com sua matriz identidade;
3. Normalizar colunas de M1; // Divide cada elemento da coluna pela soma da coluna;
4. Fazer M2 = M1 x M1 (*e* vezes) // Expande M através da sua elevação à potência *e*;
5. Fazer M3 = M2°*h* // Inflaciona M2 através do produto de Hadamard com o expoente *h*;
6. Repetir os passos 3, 4 e 5 até que M3 x M3 = M3 // A estabilidade é alcançada quando a matriz alcança a forma idempotente.

Fonte: os autores (2017).

3.3 PRÉ-PROCESSAMENTO DOS DADOS

Os dados relativos às operações realizadas na bolsa de valores contêm alta dimensionalidade, pois incluem grande quantidade de detalhes relativos a cada negócio. A aplicação do algoritmo MCL para os fins almejados nesta pesquisa não necessita da maior parte destes dados, portanto, implica na necessidade de realizar um pré-processamento visando, não

apenas excluir os dados supérfluos, mas também adequar os dados exigidos a um formato que permita construir a matriz de adjacência utilizada pelo algoritmo MCL.

Assim, os dados obtidos do Sistema de Supervisão do Mercado de Capitais Brasileiro (SSMCB v. 1.0) da CVM foram tratados de modo a organizar os negócios conforme os campos mostrados na Tabela 1, a qual consiste basicamente numa representação de um grafo em que o Vendedor e o Comprador são vértices e o Número de Vendas são as arestas. Este tratamento prévio dos dados é uma etapa normalmente trabalhosa conforme destacado por diversos autores ao apresentarem as etapas de um processo de descoberta de conhecimento em bases de dados (KDD, do original em inglês *Knowledge-Discovery in Data Bases*).

Tabela 1 - Campos dos dados de entrada.

Identificação (id) do Vendedor	**Identificação (id) do Comprador**	**N.º de Vendas**

Fonte: os autores (2017).

3.4 ESTUDO DE CASO

Em termos práticos, a manipulação de preços no mercado de capitais tem mais probabilidade de ocorrer em ativos de baixa liquidez (com poucos negócios no período). Ativos nos quais existe um grande número de investidores realizando uma grande quantidade de negócios tendem a se aproximar de mercados eficientes. Em ativos de alta liquidez, dificilmente um grupo de investidores conseguiria realizar operações que tivessem o condão de alterar significativamente os preços negociados e, mesmo se conseguisse fazê-lo, as distorções causadas seriam rapidamente corrigidas pelo mercado.

Assim, para exemplificar o uso do MCL, utilizou-se como único critério de escolha a liquidez dos ativos negociados. Com isso, optou-se pela utilização de dados dos negócios realizados com ações ordinárias da companhia Petro Rio S.A. (PRIO3), a qual apresenta relativamente poucos negócios realizados por pregão, no entanto, possuindo negócios em todos os pregões no período analisado. Os dados das operações são relativos aos dias úteis no período entre 17/out/2016 e 28/out/2016, ou seja, 10 pregões. Cabe destacar que a escolha do ativo a ser analisado e do período de análise foram arbitrárias e não existe nenhum tipo de suspeita real quanto a irregularidades nestes dados.

Neste trabalho, se optou por limitar o estudo ao mercado a vista com lotes padrão (no qual são negociadas quantidades múltiplas do lote mínimo de 100 ações), pois é desprezível a

possibilidade de irregularidades no mercado fracionário (no qual são feitos negócios com até 99 ações). Assim, no período em questão, 92 investidores negociaram com as ações PRIO3. Uma vez que esta informação não pode ser tornada pública, a identificação dos investidores foi feita através de códigos de modo a garantir o sigilo bursátil de suas operações organizadas conforme os campos da Tabela 1. Nesta tabela foram inseridos 197 registros com o número de vendas realizadas entre cada par de investidores que operou a ação no período analisado. Na Tabela 2 é apresentada uma síntese das operações realizadas pelos 92 investidores.

Tabela 2 – Número de negócios entre os mesmos investidores.

N.º de negócios entre um mesmo par de investidores	**N.º de ocorrências**	**N.º total de negócios no período**
1	158	158
2	25	50
3	11	33
4	1	4
5	2	10
6 ou +	0	0
Total	**197**	**255**

Fonte: os autores (2017).

Da Tabela 2 constata-se que, em 158 ocasiões um vendedor vendeu uma única vez para um comprador. Em 25 ocasiões, um vendedor vendeu duas vezes para um mesmo comprador e assim por diante. Destaca-se que houve apenas duas oportunidades em que um mesmo vendedor vendeu 5 vezes para um mesmo comprador, sendo este o número máximo de vezes em que um vendedor fez negócio com um mesmo comprador no período.

No entanto, além de considerar os investidores que realmente negociaram no período, foi realizada a injeção de dados de investidores e de operações fictícias dentre os dados reais visando representar a atuação em conluio de um grupo de investidores. Assim, além dos 92 investidores que realmente operaram, foram adicionados aos dados mais quatro investidores fictícios (i93, i94, i95 e i96), considerando que realizaram vendas entre si, dentro do período de análise e com a ação escolhida. Para estes investidores fictícios, foi estipulada uma dinâmica circular conforme exemplificada na atuação de v1, v2, v3 e v4 apresentados na Figura 2, exceto pelo número de vendas para uma mesma contraparte, as quais foram especificadas como 5 vendas realizadas entre três dos investidores e apenas 1 venda para um quarto investidor. A

Tabela 3 ilustra como ficaram os dados do grafo após a inserção destes quatro registros ao final do conjunto de dados. Com isso, os dados do grafo passaram de 197 para 201 registros.

Tabela 3 – Grafo com as operações fictícias inseridas ao fim dos dados reais.

Registro	Id do vendedor	Id do comprador	N.º de vendas
1	i01	i03	2
2	i01	i27	1
3	i02	i08	1
4	i03	i56	3
(...)	(...)	(...)	(...)
198	i93	i94	5
199	i94	i95	5
200	i95	i93	5
201	i95	i96	1

Fonte: os autores (2017).

Tabela 4 – Exemplo da Matriz de Adjacência.

	i01	i02	i03	i04	(...)	i93	i94	i95	i96
i01	0	0	2	0	(...)	0	0	0	0
i02	0	0	0	0	(...)	0	0	0	0
i03	0	0	0	0	(...)	0	0	0	0
i04	2	0	1	0	(...)	0	0	0	0
(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)	(...)
i93	0	0	0	0	(...)	0	5	0	0
i94	0	0	0	0	(...)	0	0	5	0
i95	0	0	0	0	(...)	5	0	0	1
i96	0	0	0	0	(...)	0	0	0	0

Fonte: os autores (2017).

Por fim, os dados foram transformados numa matriz de adjacência. Esta matriz é outra forma de representação do grafo e consiste na entrada do algoritmo. Nesta matriz, a qual está exemplificada na Tabela 4, todos os investidores são comparados com todos os investidores. A partir da primeira coluna da Tabela 4 se pode visualizar, por exemplo, que i93 realizou 5 vendas para i94; i94 realizou 5 vendas para i95; i95 realizou 5 vendas para i93; e, por fim, i95 realizou 1 venda para i96. É esta matriz que terá suas colunas reiteradamente normalizadas, expandidas e inflacionadas a partir da aplicação do MCL.

4 ANÁLISE DOS RESULTADOS

A aplicação do MCL no estudo de caso se deu através do uso da linguagem de programação R, especificamente de uma biblioteca desenvolvida por Martin L. Jäger para implementar o algoritmo (Jäger, 2015). No entanto, existem outras versões públicas do algoritmo desenvolvidas nas linguagens de programação Python e C, esta última disponibilizada pelo próprio Van Dongen (2000b). Trata-se de um ponto relevante a considerar, uma vez que as diferentes implementações do algoritmo podem sofrer adaptações e ser limitadas por restrições inerentes às linguagens de programação em que foram desenvolvidas.

Na aplicação do algoritmo em R, são disponibilizados ao usuário vários parâmetros além do expoente de expansão e de inflação. Por exemplo, é possível definir se o algoritmo pode formar agrupamentos de tamanho único (*singletons*) quando se considera que um investidor mantém uma relação pouco significativa com outros investidores ou se todos os investidores nessa condição podem ser considerados como 'ruído' e agrupados conjuntamente. Além disso, é possível determinar se a matriz idempotente do estado de equilíbrio deve ser retornada ou somente os agrupamentos. Assim, o algoritmo apresenta outras opções para o usuário além do ajuste dos parâmetros do modelo.

Em relação aos coeficientes de expansão e inflação utilizados no experimento, foi utilizado um coeficiente de expansão fixo $e = 2$, o qual é geralmente usado como valor padrão, e variou-se o coeficiente de inflação h. Para alguns dentre os diversos valores de h utilizados, o algoritmo retornou um erro na última iteração. De acordo com as explicações do implantador do algoritmo em R, nestes casos o programa não conseguiu convergir (Jäger, 2015). Constatou-se, no entanto, que a dificuldade na convergência consiste numa limitação que pode ser resultado da implementação do algoritmo, uma vez que testes utilizando os mesmos parâmetros, mas realizados com a biblioteca em Python conseguiram fazer o algoritmo convergir.

A partir da Tabela 5 constata-se que o desempenho ótimo do algoritmo se deu com $h = 1$, quando com apenas 12 iterações o algoritmo convergiu para os 8 agrupamentos identificados. Além disso, a variação do coeficiente de inflação implicou em algumas mudanças nos itens agrupados. Estas mudanças, no entanto, foram muito pequenas para os valores de h entre 0,1 e 1,0, quando poucos elementos foram mudados de grupo conforme o parâmetro foi variado. Isto já não ocorreu para $h = 1,2$, quando o número de agrupamentos dobrou e ocorreram mudanças significativas entre os elementos de cada agrupamento. Com outros valores testados para h (0,2; 0,4; 0,5; 0,7; 0,8; 0,9; 1,1; 1,3; 1,4; 1,5; ... ;2,5) não foi possível obter a convergência do algoritmo.

Tabela 5 – Resultados do algoritmo MCL para um coeficiente de expansão fixo $e = 2$ e considerando o agrupamento de vértices irrelevantes entre si.

Valor do coeficiente de inflação *h*	**N.º de iterações até a convergência**	**N.º de agrupamentos obtidos**
0,1	27	8
0,3	76	8
0,6	76	8
1,0	12	8
1,2	74	16

Fonte: os autores (2017).

O ponto a destacar no experimento é que, em todos os casos em que foi obtida a convergência do algoritmo, independentemente do valor de *h* utilizado, os investidores fictícios (i93; i94; i95 e i96) foram devidamente agrupados entre si.

É importante ressaltar que os agrupamentos não significam que de fato existem irregularidades nas operações entre os investidores agrupados. De modo geral, este não é o caso, pois há diversos motivos regulares que podem levar a uma concentração das contrapartes de um negócio. Os agrupamentos gerados pelo algoritmo devem ser entendidos apenas como indicadores que apontam para a necessidade de conferir as operações daqueles investidores perante o risco hipotético de que contenham irregularidades. Trata-se, portanto, de um mecanismo de priorização para a supervisão realizada pelos reguladores do mercado.

Assim, após a geração dos agrupamentos, se torna indispensável um importante trabalho de análise dos agrupamentos. Num primeiro momento, esta análise passa pela observação da matriz de equilíbrio obtida na convergência do algoritmo, pois os valores apresentados nesta matriz permitem visualizar a força do agrupamento e, portanto, quais agrupamentos são mais relevantes do que outros. Neste estudo de caso, o agrupamento injetado se mostrou o mais relevante em todos os experimentos realizados. A partir desta informação, a análise de potenciais irregularidades cometidas na bolsa de valores ainda é realizada pelos analistas da GMA (CVM) os quais buscam subsídios de outras fontes de informação além dos dados das operações para determinar a regularidade dos negócios.

5 CONCLUSÃO

Este artigo objetivou apresentar a aplicação do algoritmo de agrupamento MCL a partir de um estudo de caso. O trabalho foi desenvolvido no contexto da atuação supervisória da CVM sobre as operações realizadas na B3.

Após a introdução do artigo, foram apresentados maiores detalhes do contexto em que o trabalho foi desenvolvido. A importância do trabalho também foi justificada, uma vez que as análises de negócios realizadas pela CVM apresentam potenciais sérias consequências quanto à responsabilização administrativa ou penal de investigados em irregularidades cometidas no mercado de capitais. Além disso, viu-se que o contexto em questão apresenta complexidade e grande volume de dados, os quais demandam a utilização de métodos e técnicas de extração de conhecimento adequados.

Também foi apresentada a metodologia utilizada e as diversas etapas seguidas para aplicar adequadamente o MCL – desde o entendimento da representação da dinâmica de negócios na bolsa como um grafo, passando pela explicação do funcionamento do MCL e dos detalhes do estudo de caso.

Como resultado deste trabalho, constatou-se que a utilização do MCL é viável para o propósito objetivado e no contexto em que foi utilizado. O estudo de caso mostrou que os investidores associados às operações fictícias injetadas dentre os dados reais de negócios realizados na bolsa de valores foram todos devidamente agrupados. E os agrupamentos foram corretos diante de todas as variações realizadas nos parâmetros. Portanto, a utilização do MCL mostrou-se útil para a identificação de potenciais irregularidades no âmbito da supervisão realizada pela CVM, em particular pela GMA.

A partir deste trabalho e deste contexto, ainda é possível realizar diversos trabalhos futuros. Por exemplo, o experimento nesta pesquisa foi realizado com relativamente poucos dados (10 pregões) de um único ativo de baixa liquidez. Futuros trabalhos podem considerar a aplicação do MCL para centenas de pregões e de ativos, inclusive ativos que apresentam, em poucos minutos, mais negócios e investidores do que os analisados nos 10 pregões abordados neste trabalho. Nesse sentido, trabalhos futuros podem utilizar o software de agrupamento disponibilizado por Van Dongen (2000b), o qual tem a capacidade para lidar com grafos contendo milhões de vértices e arestas.

Além disso, as possíveis irregularidades no mercado de capitais não ficam restritas ao exemplo explorado neste trabalho. O MCL tem condições de agrupar investidores segundo diferentes critérios, os quais se concretizam nos possíveis tipos de relação estabelecida entre os

investidores. Assim, outra possibilidade para futuros trabalhos seria aperfeiçoar a qualificação da relação entre os investidores visando obter resultados mais significativos nos agrupamentos. Por exemplo, além de considerar o número de vendas entre dois investidores, é possível estabelecer uma métrica que combine o número de vendas com a quantidade de ações vendidas entre cada par de investidores representados no grafo ou ainda incluir nessa relação a dimensão temporal em que os negócios são realizados (maior ou menor concentração de negócios com uma mesma contraparte num maior ou menor período de tempo).

Por fim, também é possível e importante realizar trabalhos futuros envolvendo a proposta de Satuluri (2012), o qual apresentou como tese de doutorado um novo algoritmo desenvolvido a partir da evolução do MCL. Em seu trabalho, Satuluri explorou as limitações e deficiência do MCL e propôs aperfeiçoamentos que deram origem a um algoritmo de simulação multinível de fluxos estocásticos, o MLR-MCL. Portanto, é possível que o MLR-MCL possa vir a ser usado no contexto deste trabalho, com o mesmo propósito, mas com possíveis benefícios em relação ao MCL.

REFERÊNCIAS

Brasil (1976). Lei nº 6.385, de 7 de dezembro de 1976. *Diário Oficial da República Federativa do Brasil*, Poder Executivo. Brasília: DF.

Brasil (2013). *Planejamento Estratégico da CVM*: Construindo a CVM de 2023. Rio de Janeiro: RJ.

Islam, M. N., Haque, S. M. R., Alam, K.M., Tarikuzzaman, M. (2009). An approach to improve collusion set detection using MCL algorithm. *Proceedings of the 12th International Conference on Computers and Information Technology – ICCIT*, 237-242. Dhaka: Bangladesh.

Jäger, M. L. (2015). Package 'MCL' – *Reference Manual*.

Palshikar, G. K., APTE, M. M. (2005). Collusion Set Detection using Graph Clustering. *International Conference on Management of Data – COMAD*, Hyderabad: India.

Van Dongen, S. (2000a). *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht: Netherlands.

Van Dongen, S. (2000b). *MCL - a cluster algorithm for graphs*. Retrieved from https://www.micans.org/mcl/index.html

Satuluri, V. M. (2012). *Scalable Clustering of Modern Networks*. The Ohio State University: Columbus, United States of America.