

Fotografia e *Big Data*: implicações metodológicas¹

Fábio Gomes Gouveia²

Lia Scarton Carreira³

Universidade Federal do Espírito Santo, Vitória, ES

Resumo

Pesquisar grandes volumes de dados implica repensar e propor métodos diferenciados de extração, visualização e análise. O uso crescente de tecnologias digitais nas ciências humanas requer igualmente atenção aos seus processos tecnológicos e numéricos. No que tange ao estudo de imagem, apresentamos neste artigo as problemáticas metodológicas dessa forma de pesquisa a partir de iniciativas recentes, destacando a atuação do Laboratório de Estudos sobre Imagem e Cibercultura. A pesquisa por ele desenvolvida e aqui apresentada centra-se nos processos de coleta, visualização e análise de imagens acerca dos últimos protestos ocorridos no Brasil e publicadas nos sites de redes sociais Facebook e Instagram. A partir da descrição de seus processos, objetivamos compreender melhor as implicações metodológicas desse tipo de investigação científica e suas contribuições para o campo da imagem.

Palavras-chave: Fotografia; Big Data; Visualização; Metodologia.

Introdução

Nos últimos anos, diversos campos de saber têm se preocupado com regularidade com os modos de se abordar grandes volumes de dados em pesquisas científicas. Com a crescente produção e circulação de conteúdos, em especial no que tange ao digital e às redes online, tem se tornado imperativo repensar e propor métodos diversificados de coleta, análise e visualização desses dados. O que antes era terreno familiar do campo das ciências exatas passa cada vez mais a compor um campo transdisciplinar em expansão. Áreas diversas unem esforços a fim de estabelecer métodos condizentes com o atual contexto de intensa

¹ Trabalho apresentado no GP Fotografia, XIII Encontro dos Grupos de Pesquisas em Comunicação, evento componente do XXXVI Congresso Brasileiro de Ciências da Comunicação.

² Professor adjunto efetivo do Curso de Graduação em Comunicação Social da Universidade Federal do Espírito Santo (UFES). Coordenador do Laboratório de Estudos em Imagem e Cibercultura. Doutor em Comunicação e Cultura pela Universidade Federal do Rio de Janeiro. Email: fabiogv@gmail.com

³ Pesquisadora Associada do Laboratório de Estudos em Imagem e Cibercultura. Mestre em Comunicação e Cultura da Universidade Federal do Rio de Janeiro. Email: liacarreira@gmail.com

produção de conteúdos. Na comunicação, apoia-se sobre uma ampla gama de recursos teóricos e metodológicos, abrangendo também práticas da ciência da computação, dos estudos de software e da estatística.

Produz-se muito e de forma acelerada. Intensificada pelo o desenvolvimento de tecnologias da comunicação - dos dispositivos móveis aos softwares e ferramentas online -, essa produção desenfreada passou a estruturar-se enquanto objeto de pesquisas diversas. Como lidar com esse grande volume de dados? Como podemos trabalhá-los dentro das ciências humanas? Que implicações teóricas e metodológicas colocam-se em jogo? É preciso investir cada vez mais em métodos condizentes com esse contexto produtivo digital e acelerado de modo a compreender não somente seus processos, mas sua atuação enquanto objeto de investigação.

O atual cenário de produção de imagens vem sendo, igualmente, caracterizado pelo seu intenso fluxo e grande volume, pela sua distribuição global e rizomática, impulsionada pela crescente acessibilidade aos meios de produção e compartilhamento. Inseridas cada vez mais no âmbito do digital e da Internet, essas imagens são produzidas e consumidas em grande velocidade e quantidade. A *web*, como um enorme arquivo em constante expansão, tornou-se palco de uma complexa dinâmica de trocas, cujos rastros podem ser identificados, destacados e analisados. O caráter numérico de seus documentos, extraídos dos mais diversos contextos, atuam na *web* como vestígios das atividades em rede.

À exemplo dos últimos protestos ocorridos no mês de junho em todo Brasil, essa produção incessante de imagens atrelada ao potencial da distribuição em rede toma proporções não somente estéticas, mas políticas. Uma câmera na mão interconectada a milhares de usuários em rede mostrou-se mais do que nunca uma potência criativa e crítica, que une a atividade online à mobilização de rua. Encontrar modos de trabalhar com esses dados e seus rastros é, portanto, fundamental para a compreensão não somente dos modos contemporâneos de produção de imagens, mas também dos aspectos políticos e sociais dessa prática de produção e compartilhamento que se tornou cotidiana.

Neste artigo, buscamos, portanto, apresentar algumas das características e problemáticas desses processos de investigação científica de grandes volumes de dados (em especial ao

que tange os estudos da imagem), ressaltando sua importância no âmbito das ciências humanas e destacando algumas iniciativas já em desenvolvimento. Em um primeiro momento, discorreremos brevemente sobre as implicações teóricas e metodológicas desses processos no que se convencionou chamar de *digital humanities*, para em seguida apresentar dois exemplos de pesquisas: uma internacional, desenvolvida pelo Software Studies Initiative, e uma nacional, do Laboratório de Estudos sobre Imagem e Cibercultura da Universidade Federal do Espírito Santo, do qual fazemos parte.

Questões metodológicas dos processos de investigação de grandes dados

A expressão “imagem digital”, que em francês traduz-se para “*image numérique*”, nunca foi tão propícia para caracterizar o nosso atual contexto. O caráter numérico de qualquer arquivo digital – incluindo as fotografias - é justamente o que permite que hoje sejam desenvolvidas análises de gigantescos conjuntos de dados – os chamados *Big Data*. O termo, que adquire destaque cada vez maior na mídia e no meio acadêmico, designa um grande volume de dados que requer não somente tecnologias determinadas (como dispositivos com maior capacidade de processamento e armazenamento, softwares de extração e visualização, etc.), mas também métodos e profissionais especializados.

Trabalhar com esse grande volume de dados envolve, portanto, três etapas fundamentais, cujas bases estão em seu caráter digital: envolve os processos de extração, de visualização e de análise dos dados. Sua composição numérica possibilita, desse modo, a coleta de informações-chaves para análise. Ela permite seu armazenamento computacional e possibilita trabalhá-los através do uso de algoritmos determinados. Os dados coletados podem ainda ser recombinaados, correlacionados, contabilizados e classificados. No processo de extração de publicações de um site de rede social, por exemplo, podemos recolher não somente a parte textual, como a localização geográfica, a data e a hora exata de sua postagem. Da mesma forma, é possível identificar relações que estabelecem com outras postagens e, conseqüentemente, com outros usuários dessa mesma ferramenta *online*. Esses dados podem, ainda, ser visualizados de modos diferentes: podemos “traduzir” os textos coletados em dados numéricos, por exemplo, compondo uma espécie de infográfico; podemos analisar cada palavra publicada e evidenciar suas relações e proporções em rede;

podemos compor quadros relacionais e interativos, nos quais pode-se observar semelhanças e dissonâncias entre as publicações, entre os nós de uma rede ou entre os dados computados, como sua geolocalização⁴. Há, portanto, uma multiplicidade de modos de exposição desses dados coletados e computados. À cada modo de visualização, novos padrões e divergências podem ser observadas.

É, portanto, o próprio caráter digital do arquivo que constituirá a base dos processos subsequentes de investigação de grandes dados. O documento que se deseja investigar deve, nesse sentido, passar por essa reconstituição numérica para que possa ser computado, correlacionado e mapeado para visualização e análise. Mas engana-se aquele que acredita que a base numérica determinará a análise como somente quantitativa. É essa a uma das grandes questões que vem sendo discutida acerca das pesquisas de grandes dados.

Para pesquisadores como David Berry (2011), acredita-se que a tecnologia digital está transformando fundamentalmente os modos pelos quais realizamos nossas pesquisas. Em *The Computational Turn* (2011), Berry argumenta que cada vez mais as investigações científicas, das áreas mais diversas, voltam-se para as tecnologias digitais, modificando o que se compreende por pesquisa científica. Essa transformação pode ser observada nos próprios modos de busca bibliográfica ou de coleta de dados, por exemplo. Observa-se também uma crescente necessidade de digitalização de material de análise, como a produção e reprodução de imagens de arquivos físicos (sejam eles fotografias, livros de uma biblioteca, obras de arte ou até mesmo objetos de investigação das áreas de biologia, química, etc.).

É certo, contudo, que o nível e a abrangência dessa transformação variam de área para área. E que essa expansão do uso de tecnologias digitais vai além das disciplinas tradicionais como a ciência da computação e da estatística, mas não se constitui como fator hegemônico. Entretanto, observa-se uma transformação aparente na constituição geral do que se define principalmente por ciência humanas. A própria proposição de uma *Digital Humanities* (que pode ser traduzido por Ciências Humanas Digitais), implica essa mudança conceitual. O termo, que antes se designava por “*computing in the humanities*” (computação nas ciências humanas), como descreveu Berry (2011), passou a caracterizar

⁴ Ver Malini (2013) para pesquisas recentes sobre coleta de publicações no Twitter, como exemplo.

um campo de estudo que ultrapassa o mero uso da computação como auxiliar nas pesquisas das ciências humanas. A computação e o estudo de software passaram, nas últimas décadas, a realizar um papel central no desenvolvimento dessas pesquisas. E isso não implica, como ressalta o próprio Berry, numa “quantificação” total das ciências humanas. Pelo contrário, o uso de tecnologias digitais contribui também para a compreensão de fenômenos sociais e culturais complexos, que fogem da apresentação meramente numérica e estatística. O uso crescente das tecnologias digitais contribui, portanto, para uma outra conceituação de ciência (e, como apresentado por Berry, para uma outra noção de “universidade” enquanto instituição de ensino e pesquisa). Trata-se, dessa forma, de uma conceituação que dilui as fronteiras entre disciplinas, e seus processos metodológicos característicos, direcionando-se para um campo cada vez mais transdisciplinar.

Nesse sentido, as pesquisas contemporâneas que envolvem o uso de tecnologias digitais, devem se atentar não somente às suas especificidades teóricas e práticas, mas compreender também esses processos tecnológicos-digitais. Ou seja, no que diz respeito ao estudo de imagem (ainda que atualmente seja complexo definir bem essas áreas e disciplinas), devemos compreender também as implicações de sua composição numérica, as bases de seus softwares, os sites de redes sociais dos quais são extraídos, por exemplo, e etc. Em se tratando de pesquisas que envolvem grandes quantidades de imagens, e que por conseguinte necessitam de tecnologias de alto desempenho, de grande capacidade de armazenamento, de softwares especificamente criados para extração e visualização, é preciso cada vez mais compreender suas especificidades técnicas.

Ao mesmo tempo, é preciso aliar essa abrangência tecnológica no campo das ciências humanas aos estudos estéticos, sócio, político e culturais, como fazem laboratórios de pesquisa como o Software Studies Initiative nos Estados Unidos. Coordenador pelo pesquisador Lev Manovich, este laboratório voltado para o estudo de software realiza coleta, análise e visualizações de imagens desde 2007. O laboratório já conta com uma ampla utilização dos processos e variedade de imagens, a ponto de propor seus próprios aplicativos e de ampliar cada vez mais os limites de extração de dados. Seu foco está, portanto, tanto no processo de desenvolvimento e estudo de softwares como na análise estética das imagens.

Em 2009, o laboratório trabalhou com mais de 1 milhão de imagens de *mangás* digitalizados e coletadas a partir de páginas de *upload* e de tradução de fãs (*scanlation sites*), de modo a identificar e analisar seus aspectos visuais característicos. Em outras ocasiões, o laboratório comparou pinturas de Piet Mondrian e Mark Rothko (para destacar um exemplo cujas fontes não advêm somente de conteúdos gerados *online*), cujas visualizações nos possibilitam traçar semelhanças e diferenças entre as obras ao longo do tempo, destacando não somente aspectos físicos, mas aspectos característicos de uma produção artística de um tempo específico. Outras abordagens envolvem ainda imagens de filmes, como os de Dziga Vertov, de *videogames*, de capas de revistas e de publicações em sites de redes sociais, que abrangem noções técnicas dessas imagens e também sociais e históricas.

Sob essa mesma perspectiva, o Laboratório de Estudos sobre Imagem e Cibercultura (Labic) da Universidade Federal do Espírito Santo passou a realizar em 2012 um estudo sobre imagens coletadas em sites de redes sociais. Inserida no projeto Visagem, procuramos investigar nesta pesquisa as imagens publicadas pelos usuários de dois sites de redes sociais na internet: Instagram e Facebook. Em se tratando de imagens produzidas no contexto dos recentes protestos ocorrido no mês de junho de 2013 no Brasil, esta pesquisa visa não somente compreender esses processos de produção de imagem contemporâneos, mas também as especificidades desses movimentos sociais. Contudo, devido ao seu caráter embrionário, apresentaremos neste artigo as considerações iniciais sobre os processos de extração e visualização dessas imagens, de modo a compreendermos melhor as questões metodológicas desse tipo de investigação. O objetivo inicial foi, portanto, verificar as diferenças e aproximações entre os dois sites no que tange às imagens dos usuários.

Imagens do #protestoes no Instagram e no Facebook: diferenças visuais

O Laboratório de Estudos sobre Imagem e Cibercultura (Labic) já desenvolve há alguns anos pesquisa voltadas para grandes volumes de dados. Contudo, estas centram-se em dados textuais coletados online. Recentemente, o laboratório voltou-se para conteúdos publicados sobre o Marco Civil da Internet e sobre os últimos protestos acerca do transporte público brasileiro, gerados em sites de redes sociais como o Twitter. Com os dados

extraídos dessa ferramenta *online*, pode-se visualizar, a partir de *softwares* determinados, traços estabelecidos em rede que podem servir para caracterizar esses movimentos.

Da mesma forma, o laboratório passou recentemente a voltar-se também para as imagens publicadas em redes. Apresentaremos aqui alguns aspectos metodológicos sobre essa pesquisa, cujo foco está nas imagens acerca dos últimos protestos de junho de 2013 no estado do Espírito Santo.

A pesquisa usou como método de coleta a extração manual das imagens, que ficaram armazenadas em disco rígido para posterior análise. De cada site de relacionamento foram extraídas cerca de 500 imagens, que compreendiam entre elas fotografias, montagens, cartazes, ilustrações, convocações. Com estas imagens, foram coletados ainda outros dados acerca dos perfis (autores) das publicações, assim como dos endereços *online*, do tipo de documento e da data de postagem na rede.

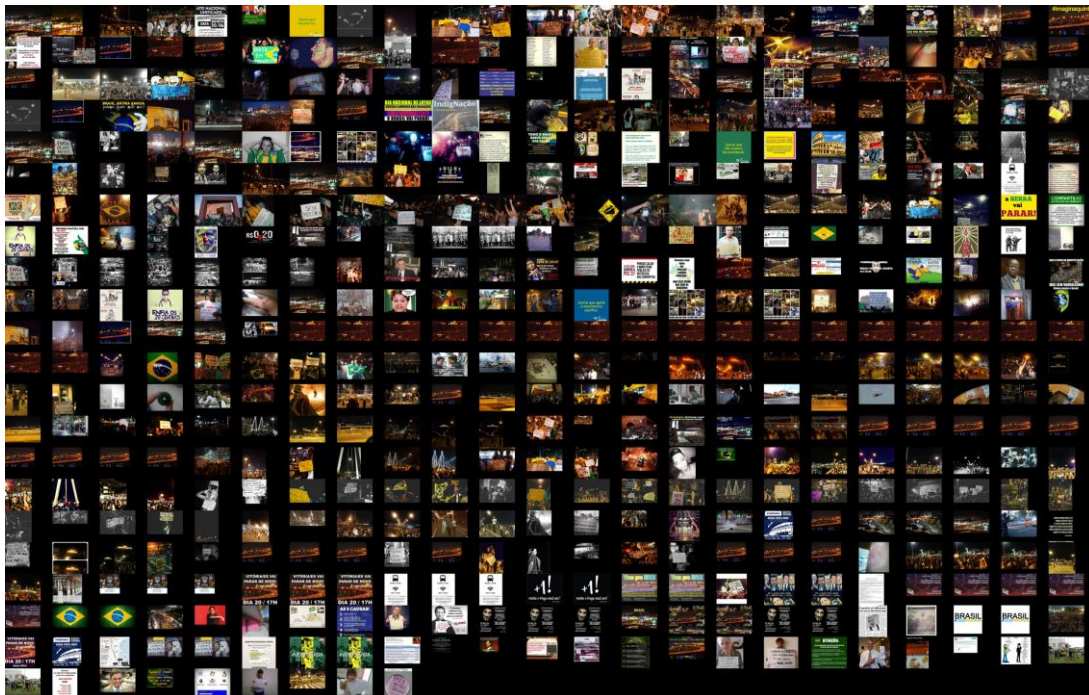


Imagem 01: mosaico criado a partir de imagens coletadas do Facebook.

Os processos de extração de cada site de rede social são, contudo diferentes. Cada ferramenta de relacionamento online possui características próprias, não somente de exibição, mas também de compartilhamento. Para a extração de imagens do Facebook foram utilizados os seguintes procedimentos: a digitação da palavra-chave ou hashtag

“protestoes”⁵ no espaço para pesquisa; a captura manual de todas as imagens que surgiram a partir da busca; arquivamento das imagens em disco rígido do computador para computar os dados; e registro dos dados de cada imagem em uma tabela digital com as informações do perfil de que publicou a imagem, da categoria da imagem, e da data de sua publicação.

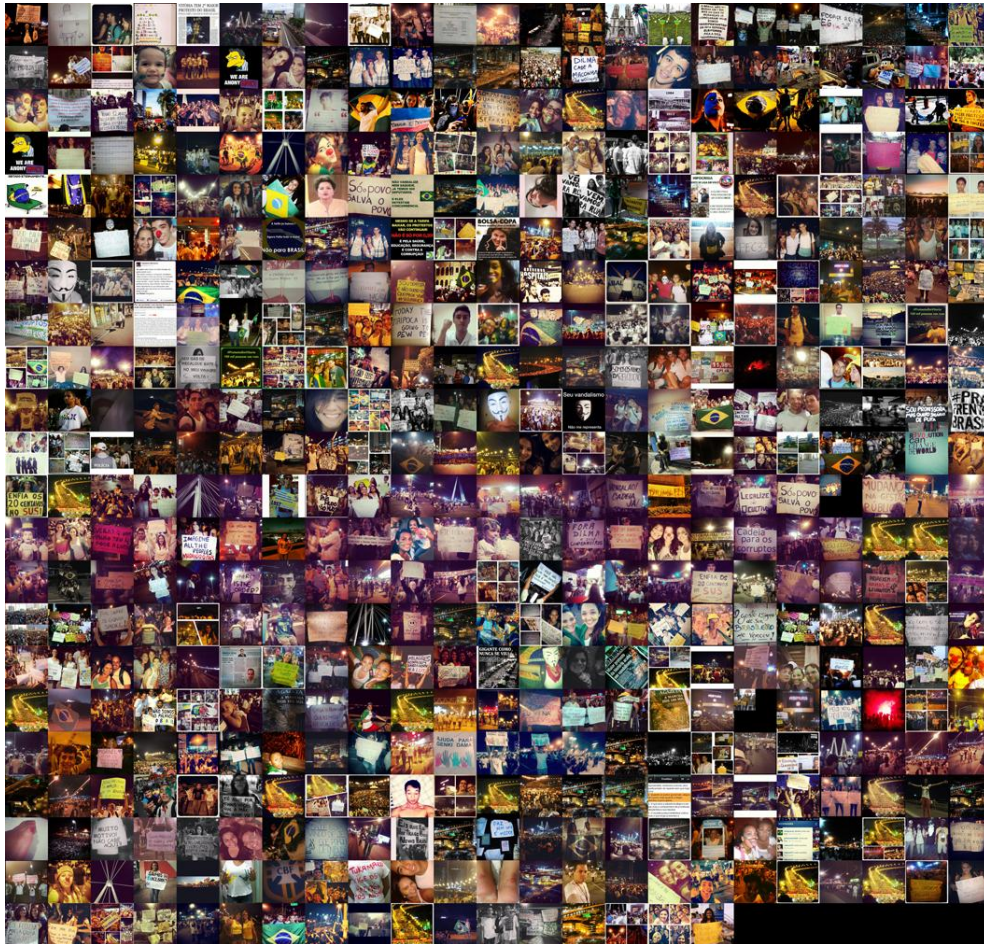
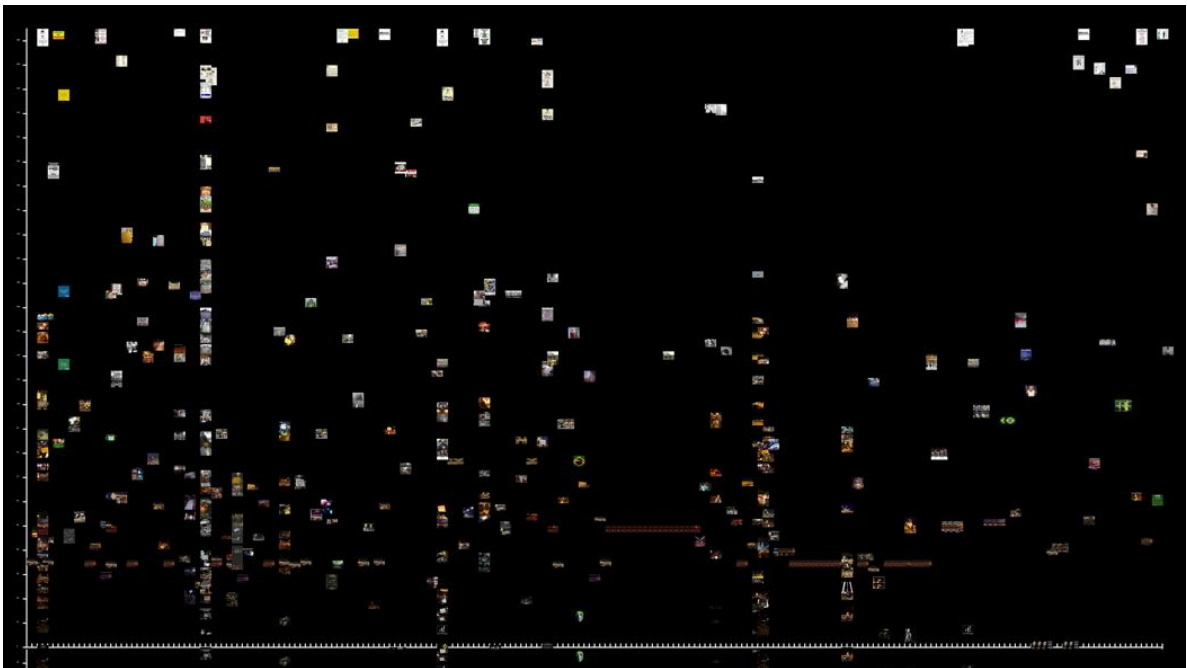


Imagem 02: mosaico criado a partir de imagens coletadas do Instagram.

Já na extração das imagens da rede social Instagram, usamos o site Webstagram, que ordena e permite que as pessoas acessem imagens do Instagram por meio de pesquisa com hashtags ou palavras específicas. Assim, fizemos a coleta de dados manualmente a partir do dia 24 de junho, somando quase três mil imagens com a hashtag #protestoes. Como o intuito era fazer a comparação com as imagens do Facebook, analisamos neste trabalho apenas as primeiras 500 imagens extraídas, de modo a elaborarmos um panorama semelhante entre as duas redes sociais.

⁵ O uso do símbolo “#” refere-se ao termo “hashtag”, que representa palavra-chave, mas também isola a expressão do texto corrido. Para capturarmos o perfil, o autor da postagem precisou inserir a expressão “#protestoes” em sua mensagem. Nossa busca ignorou a presença ou ausência das palavras “protesto” ou “protestoes” no texto.

Tendo as imagens arquivadas e catalogadas, procedeu-se o processamento no programa ImageJ⁶, usando a macro ImagePlot⁷, que resultou em diversas possibilidades de visualização de ambos conjuntos de imagens coletadas. A primeira delas foi uma comparação dos mosaicos dos dois conjuntos (Imagem 01 e 02). Visualmente podemos destacar as variações tonais de todas as imagens do Facebook e do Instagram. No Instagram há mais variações cromáticas, sendo que isso deve-se em grande parte em função dos filtros disponíveis no aplicativo. Já nas imagens do Facebook, há nítida separação de tons, que evidenciam a natureza mais “crua” das imagens, uma vez que a publicação das imagens nesta rede social é feita normalmente de modo direto, sem intervenção de filtros, as cenas exibidas têm uma coloração com menor valor de brilho.



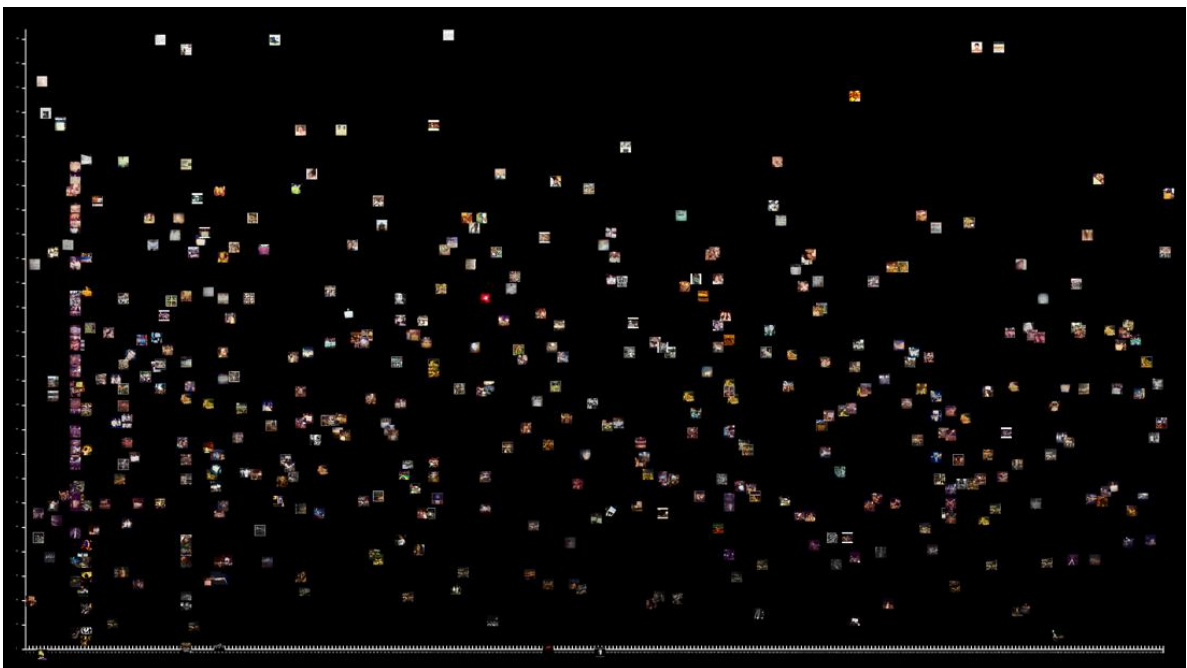
Imagens 03: gráficos de brilho médio x perfis de imagens do Facebook.

Uma primeira comparação permite, então, visualizar o nível de intervenção dos perfis que publicaram as imagens em cada uma das duas redes: há mais modificações com filtros nas

⁶ “O ImageJ é um software para processamento e análise de imagens, desenvolvido por Wayne Rasband no National Institute of Mental Health, USA, em linguagem Java. Com este software é possível exibir, editar, analisar, processar, salvar e imprimir imagens de 8, 16 e 32 bits. Permite o processamento de diversos formatos de imagem como TIFF, GIF, JPEG, BMP, DICOM e FITS.” (HANNICKEL et. al., 2012, 17).

⁷ ImagePlot é um software livre de ferramentas, desenvolvido pela equipe do professor Lev Manovich, do Centro de Pesquisa *Software Studies*, para visualizar coleções de imagens ou vídeos de qualquer tamanho. Ele é implementado como uma macro que trabalha com o programa de processamento de imagem ImageJ. Veja mais em (<http://lab.softwarestudies.com/p/imageplot.html>)

imagens do Instagram, inclusive há uma predominância do filtro que causa efeito de “envelhecimento” da imagem, como se as fotos tivessem passado por um processo de deteriorização. Ainda que os dois conjuntos de imagens tenham sido produzidos em situações noturnas, as fotos publicadas no Facebook contêm mais áreas de baixas luzes, enquanto que as cenas do Instagram possuem mais tons claros. Este fato se explica por duas razões: as fotos do Instagram contêm enquadramento com pessoas em primeiro plano e alguns cartazes, o que torna a imagem mais clara; e as fotos do Facebook têm em, sua maioria, planos mais abertos (o que contribuiu para termos cenas mais escuras).



Imagens 04: gráficos de brilho médio x perfis de imagens do Instagram.

Outra análise possível dos dois conjuntos de imagens é a partir de um gráfico que dispõe as informações de brilho médio no eixo Y e os perfis que publicaram as imagens no eixo X (Imagens 03 e 04). Deste modo observamos, entre outras informações, como as diferenças nos valores de brilho permitem separar os conjuntos de imagens em sub-grupos. Os tons médios predominantes nas imagens do Instagram fizeram com que as imagens aparecessem mais distribuídas no gráfico, ao contrário do que ocorre com o Facebook, que concentra as imagens em grupos de altas e baixas luzes. O sub-grupo com pouco brilho médio representa as fotos feitas com condição precária de iluminação, ou seja, a maioria. Mas há ainda um grande número de imagens com alto brilho médio. Estas são, no Facebook, cartazes, convocações, infográficos, ou análises, que foram publicadas como imagens. Já no

Instagram esse tipo praticamente não aparece. Ainda que não sejam fotografias de um acontecimento, mas sim de um objeto, a existência desse grupo de imagens é reveladora do modo de utilização das redes sociais. Se no Facebook há a multiplicidade de tipos de imagens, no Instagram predomina-se um tipo de fotografia. O caráter aberto e fechado - Facebook e Instagram, respectivamente - se reflete no tipo de imagem publicada.

Por fim, o terceiro viés de análise das imagens deste artigo se detém nos grupos de imagens que mais foram publicadas e nos perfis mais ativos. Percebe-se que cinco perfis foram os que mais inseriram fotos em suas linhas do tempo no Facebook, enquanto que no Instagram há apenas um perfil de destaque. Além disso, é possível inferir ainda as principais imagens dos dois conjuntos. Aquelas que foram muito compartilhadas aparecerão alinhadas horizontalmente. Aí temos um fato de distinção entre as duas redes sociais. Enquanto que no Facebook o mecanismo de compartilhamento simples cria um efeito de difusão e de consolidação de imagens-ícones, no Instagram isso praticamente não ocorre. Em virtude disso, o Instagram apresenta suas imagens pulverizadas. O Facebook, por sua vez, mostra algumas fotografias que passaram a ocupar o imaginário local como emblemáticas: a principal ponte do Espírito Santo completamente tomada por manifestantes. A ocupação da Terceira Ponte, que não possui permissão para passagem de pedestres, foi emblemática principalmente no dia 17 de junho, quando pela primeira vez a travessia aconteceu. O fato seria repetido no dia 20 de junho, mas sem os letreiros informativos com os dizeres “Ponte Interditada” e “Por Manifestantes”. A reprodução de imagens mostrando essas duas expressões demonstra o peso destas cenas.

Considerações finais

A base de nossas análises está no caráter técnico da imagem digital. Se o próprio conceito de Imagem Técnica (Flusser, 2002) já era de maneira muito consistente utilizada nos estudos de fotografia analógica, a essência numérica da imagem e dos arquivos digitais permite decompor a informação visual em sequências de números. Isso possibilita investigar uma amplitude muito grande de imagens de uma forma que jamais foi feito na história da fotografia. A decomposição da imagem em número para possibilitar o processamento dos dados e a recomposição da fotografia para que possa ser analisada

dentro do conjunto de imagens, torna possível identificar padrões, agrupar elementos, identificar produtividade de autores, entre outras tantas análises possíveis.

A utilização de novas ferramentas de visualização de grandes conjuntos de dados abre ao campo de análise de imagem imensos desafios e perspectivas promissoras. Para além de uma investigação dos elementos binários de cada fotografia, ou seja, cada pixel individualmente, estamos diante de uma ruptura com a leitura que universaliza o pontual. Com esses mecanismos, alguns pesquisadores já começam a vislumbrar a possibilidade de comparar todas as imagens da História da Fotografia.

Contudo, separar as informações de cada pixel ou os metadados da fotografia por si só não serve como conclusão definitiva. É possível comparar e mapear o comportamento dos fotógrafos por meio dessa cartografia imagética. Mas não basta. As questões que surgem com a fotografia digital estão apenas começando. Não é a quantidade de brilho de cada imagem que importa, mas sim as relações destas imagens com o todo.

Referências bibliográficas

BERRY, David. The Computational Turn: Thinking About the Digital Humanities. *Culture Machine*. v. 12, 2011. Disponível em: <<http://www.culturemachine.net/index.php/cm/article/view/440/470>>. Acesso em: 10 jul 2013

FLUSSER, Vilém. *Filosofia da Caixa Preta Ensaio para uma futura filosofia da fotografia*. Rio de Janeiro: Relume Dumará. São Paulo: Hucitec, 2002.

HANNICKEL, Adriana, et. al. ImageJ como ferramenta para medida da área de partículas de magnetita em três escalas nanométricas. In: *Revista Militar de Ciência e Tecnologia*, Vol. XXIX, 4º. Trimestre 2012. Artigo on line disponível em http://rmct.ime.eb.br/arquivos/RMCT_4_tri_2012/RMCT_057_E4B_11.pdf

MALINI, Fábio. A batalha do vinagre: por que o #protestoSP não teve uma, mas muitas hashtags. Laboratório de Estudos sobre Imagem e Cibercultura. Publicado em: 14 jun 2013. Disponível em: <<http://www.labic.net/cartografia-das-controversias/a-batalha-do-vinagre-por-que-o-protestosp-nao-teve-uma-mas-muitas-hashtags/>>. Acesso em: 24 jun 2013