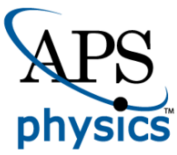

Nondeterministic Approach to Tree-Based Jet Substructure

Stephen D. Ellis, Andrew Hornig, Tuhin S. Roy, David Krohn, and Matthew D. Schwartz

Phys. Rev. Lett. **108**, 182003 — Published 4 May 2012

DOI: 10.1103/PhysRevLett.108.182003

Q-jets: A Non-Deterministic Approach to Tree-Based Jet Substructure

Stephen D. Ellis,* Andrew Hornig,† and Tuhin S. Roy‡
Department of Physics, University of Washington, Seattle WA, 98195

David Krohn§ and Matthew D. Schwartz¶
Department of Physics, Harvard University, Cambridge MA, 02138
(Dated: March 19, 2012)

Jet substructure is typically studied using clustering algorithms, such as k_T, which arrange the jets' constituents into trees. Instead of considering a single tree per jet, we propose that multiple trees should be considered, weighted by an appropriate metric. Then each jet in each event produces a distribution for an observable, rather than a single value. Advantages of this approach include: 1) observables have significantly increased statistical stability; and, 2) new observables, such as the variance of the distribution, provide new handles for signal and background discrimination. For example, we find that employing a set of trees substantially reduces the observed fluctuations in the pruned mass distribution, enhancing the likelihood of new particle discovery for a given integrated luminosity. Furthermore, the resulting pruned mass distributions for (background) QCD jets are found to be substantially wider than that for (signal) jets with intrinsic mass scales, e.g. boosted W jets. A cut on this width yields a substantial enhancement in significance relative to a cut on the standard pruned jet mass alone. In particular the luminosity needed for a given significance requirement decreases by a factor of two relative to standard pruning.

To develop intuition about high-energy collisions like those at the LHC it is often helpful to think of an event as being produced by a multi-stage process. In this picture, a short distance scattering produces a few hard partons. The partons then shower soft and collinear QCD radiation. Finally, at long distances, the (colored) partons bind into the (color singlet) hadrons that we observe in the detector. This parton-shower picture explains how clusters of nearby final-state particles, called jets, defined by a jet algorithm, can reveal something about the short-distance physics. Simulations of the parton shower produce events which, with sufficient tuning, exhibit remarkable agreement with collider data for nearly any conceivable infrared safe observable.

If one takes the parton-shower picture literally, the constituents of a jet arise from a shower-like series of $1 \to 2$ splittings producing a "tree" structure. Since the shower model for QCD is dominated by soft and collinear splittings, any deviation from this behavior could indicate the presence of contamination within the jet, or might indicate that the jet is not purely of QCD origin (e.g., it could come from a boosted heavy particle). Thus, by associating trees (by "trees," we mean "clustering histories") to jets one can obtain useful information, and indeed this is the basis for much of the work in the field of jet substructure (see Ref. [1] for a review).

The association of a tree to a jet naturally emerges from the parton-shower picture. In the parton shower, soft and collinear radiation is emitted in a particular sequence: a p_{T}-ordered shower builds a tree by adding on emissions in decreasing order of transverse momentum, while an angular ordered shower adds emissions in a sequence of decreasing angle. The recombination jet algorithms try to match this behavior. The $\mathrm{k_T}$ algorithm [2] assembles a jet in increasing order of the (relative) $\mathrm{k_T}$ metric that depends on both angle and the magnitude of the momentum, and the Cambridge/Aachen (C/A) algorithm [3] assembles in increasing order of angle. Both can be viewed as a reasonable guess for the showering sequence history.

One problem with thinking of jet algorithms as reversing the parton shower is that the parton shower is not invertible – a given set of four-momenta of final state particles could have evolved through a multitude of intermediate trees. In this paper we propose a way to account for the non-invertible nature of the parton shower by associating to each jet a set of trees instead of a single tree.

Related ideas have been discussed in the past. Long ago a probabilistic approach was used to improve the behavior of seeded jet algorithms [4]. More recently, it has been shown that combining even highly correlated observables, such as jet masses arising from different grooming techniques [5], can improve discovery significance. In addition, Ref. [6] considered associating multiple trees to a jet to compare with models of showering in signal and background processes, and Ref. [7] proposed a measure of jet *fuzziness* to gauge the ambiguity in jet reconstruction. However, our approach is fundamentally different from these previous studies. We are interested in observables constructed from a *distribution* of trees for each jet in each event. For instance, we will show that by averaging tree-based observables over the trees for each jet, their statistical stability can be substantially improved.

Associating a set of trees to a jet would not be feasible if one had to consider *every* tree which could be formed from a given set of final state four-momenta in a jet. Fortunately, good approximations to such distributions obtained using every tree can be captured through a procedure analogous to Monte-Carlo integration, allow-

ing us to use a very small fraction of the trees. This is possible beause infrared and collinear safe jet observables must be insensitive to small reshufflings of the momenta, implying that large classes of trees give very similar information.

The algorithm we propose assembles a tree via a series of $2 \to 1$ mergings:

1. At every stage of clustering, a set of weights ω_{ij} for all pairs $\langle ij \rangle$ of the four-vectors is computed, and a probability $\Omega_{ij} = \omega_{ij}/N$, where $N = \sum_{\langle ij \rangle} \omega_{ij}$, is assigned to each pair.

2. A random number is generated and used to choose a pair $\langle ij \rangle$ with probability Ω_{ij}. The chosen pair is merged, and the procedure is repeated until all particles all clustered.

This algorithm directly produces trees distributed according to their weight $\prod_{\text{mergings}} \Omega_{ij}$. To produce a distribution of trees for each jet, this algorithm is simply repeated N_{tree} times (not necessarily yielding N_{tree} *distinct* trees). Note that any algorithm which modifies a tree during its construction (e.g., jet pruning) can be adapted to work with this procedure, as demonstrated below.

One particularly interesting class of weights is given by

$$\omega_{ij}^{(\alpha)} \equiv \exp\left\{-\alpha \frac{(d_{ij} - d^{\min})}{d^{\min}}\right\}. \tag{1}$$

with α a real number we call *rigidity*. Here, d_{ij} is the jet distance measure for the $\langle ij \rangle$ pair, e.g.,

$$d_{ij} = \begin{cases} d_{\text{k}_\text{T}} \equiv \min\{p_{Ti}^2, p_{Tj}^2\}\Delta R_{ij}^2 \\ d_{\text{C/A}} \equiv \Delta R_{ij}^2 \end{cases}, \tag{2}$$

where $\Delta R_{ij}^2 = \Delta y_{ij}^2 + \Delta\phi_{ij}^2$, and $d^{\min}$ is the minimum over all pairs at this stage in the clustering. Note that with this metric, our algorithm reduces to a traditional clustering algorithm when $\alpha \to \infty$, i.e., in that limit the *minimal* d_{ij} is always chosen. In this sense, it is helpful to think of the traditional, single tree algorithm as the "classical" approach, with $\alpha \sim 1/\hbar$ controlling the deviation from the "classical" clustering behavior. With this analogy, we call the trees constructed in this non-deterministic fashion Q-jets ("quantum" jets).

In order to get the most information out of the Q-jets, it is logical to consider observables which are sensitive to the ordering of the clusterings in the tree. One such observable is the pruned jet mass, which we will use as our illustrative example. As described in Ref. [8] pruning is one of the jet grooming tools [9]. It is used to sharpen signal and reduce background when considering boosted heavy objects. The basic idea is to move along the tree and try to discard radiation which is soft and not collinear, and therefore likely to represent contamination

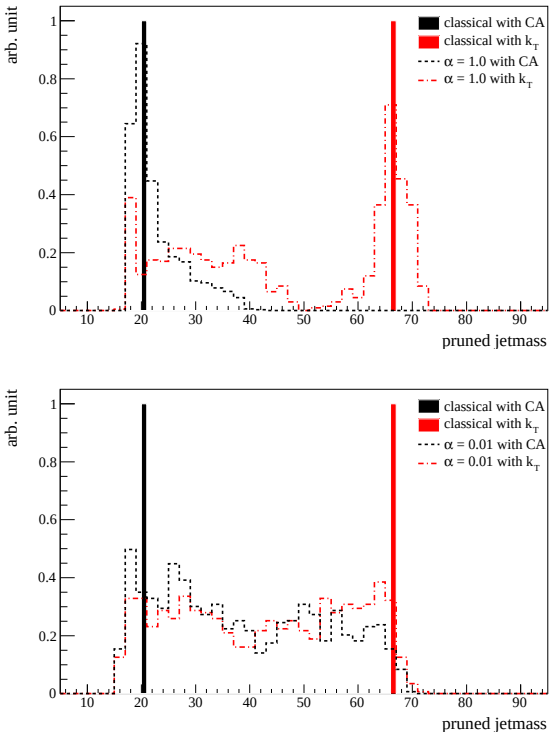

FIG. 1. Distribution of pruned jet mass for a single QCD-jet with $p_T \sim 500$ GeV. The black and red solid lines show the classical pruned masses when C/A and k_T algorithms are used to cluster the jet. The black and dashed (red and dot-dashed) line shows the pruned jet mass distribution of 1000 trees (constructed from the same jet in the same event), when the C/A (k_T) measure is used in Eq. (1). These distributions result from clusterings with rigidity $\alpha = 1.0$ (top) and $\alpha = 0.01$ (bottom).

from a part of the event in which we are not particularly interested (like the underlying event). In detail, if a step in the clustering would merge particles i and j which satisfy

$$z_{ij} \equiv \frac{\min\left(p_{T_i}, p_{T_j}\right)}{|\vec{p_{T_i}} + \vec{p_{T_j}}|} < z_{\text{cut}} \qquad \text{and} \qquad \Delta R_{ij} > D_{\text{cut}}\,, \tag{3}$$

then the merging is vetoed and the softer of the two four-momenta is discarded. In the specific analysis described here we take $z_{\text{cut}} = 0.1$ and $D_{\text{cut}} = m_{\text{jet}}/p_{\text{jet}}$, which are typical cuts for the C/A algorithm.

We apply this pruned Q-jets procedure to samples of simulated boosted W (signal) and QCD (background) jets generated with `Pythia v6.422` [10] with p_T-ordered showers using the Perugia 2011 tunes [11] and assuming a 7 TeV LHC. In lieu of detector simulation we group the visible output of Pythia into massless $\Delta\eta \times \Delta\phi = 0.1 \times 0.1$ "calorimeter cells" (with $|\eta| < 5$), preserving the energy and the direction to the cell. The cells with energy bigger than 0.5 GeV become the inputs to the initial jet-

finding algorithm (small alterations to this cut have no appreciable impact on our results). To find the initial jets we use the anti-k_T algorithm [12] with $R = 0.7$ as implemented in `Fastjet v2.4.2` [13] and require $p_T^{\rm jet} \geq 500$ GeV. Once a jet is identified, the cells clustered in the jet become input to the Qjet-pruning algorithm. A fastjet plugin with this implementation of Q-jets is available at `http://jets.physics.harvard.edu/Qjets`.

Consider first a single QCD jet from the sample described above. Fig. 1 exhibits the pruned mass distribution for this jet obtained with the classical procedure for both k_T and C/A pruning (the 2 vertical lines) and with $N_{\rm tree} = 1000$ using both the k_T and C/A metrics for d_{ij} in Eq. (1). The curves illustrate the dependence on the form of d_{ij}, as well as on the value of the rigidity parameter α. The upper panel is for $\alpha = 1.0$ where the trees are confined to stay close to the classical tree and the pruned masses likewise stay near the corresponding classical result. For small enough α (say, $\alpha \lesssim 0.1$), a broad spectrum of trees is sampled. This is shown in the lower panel of Fig. 1 for $\alpha = 0.01$, where the distributions generated with the k_T and C/A definitions of the distance d_{ij} look similar, and have little correspondence with the classical results. This suggests that for a small enough rigidity parameter pruned Q-jets become independent of the choice of distance measure used; they are therefore more likely to be characterizing physical features of an event rather than artifacts of using a particular jet algorithm.

We will now discuss two fundamentally different ways in which the discovery potential (e.g.for finding boosted W jets on top of their QCD background) can be enhanced using Q-jets:

- Observables have smaller statistical variation. Even for the same number of background jets, the use of Q-jets reduces the background fluctuations δB and increases the discovery potential $S/\delta B$, where S and B are the numbers of signal and background jets in the signal window and δB denotes the fluctuation in B.

- Qualitatively new observables, which depend on there being a distribution of trees for each jet, can now be considered. For example, we define below a powerful observable we call *volatility* which measures the width of the pruned Q-jet mass distribution for each jet, something inaccessible to a classical jet algorithm

To quantify the first of these points, we consider a large number of pseudo-experiments, each of which analyses N_J jets, with N_J taken from a Poisson distribution with mean $\langle N_J \rangle$. With a classical jet algorithm we can extract a significance by counting, in each pseudo-experiment, the number S and B, of W jets or QCD jets respectively, with pruned mass in a signal window, say between $70 - 90$ GeV. The significance is then given by $\langle S \rangle / \delta B$, where $\langle S \rangle$ is the average over the pseudo-experiments of the number of signal events in the window and δB is the RMS fluctuation of B over those pseudo-experiments. As expected $\langle S \rangle$ and $\langle B \rangle$ are proportional to $\langle N_J \rangle$, while δS and δB vary with $\sqrt{\langle N_J \rangle}$. In addition to looking at $\langle S \rangle / \delta B$, we can also look at the RMS fluctuations in the average pruned Q-jet mass of the signal jets, $\delta \langle m \rangle$, averaged over the signal jets in the signal window for each pseudo-experiment. This tells us the statistical uncertainty with which the W mass could be measured from these events.

	Vol. cut ($\mathcal{V}_{\rm cut}$)	Rigidity $\alpha = 0$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 100$
$\frac{\langle S \rangle / \delta B\|_{\rm Q}}{\langle S \rangle / \delta B\|_{\rm cl}}$	**None**	1.07(1)	1.13(1)	1.18(1)	1.14(1)	1.06(1)
	0.05	1.43(4)	1.44(3)	1.39(3)	1.27(1)	1.08(1)
	0.04	1.51(4)	1.45(4)	1.39(3)	1.29(3)	1.10(1)
	0.03	1.51(2)	1.45(3)	1.37(4)	1.35(2)	1.10(1)
	0.02	1.28(5)	1.24(3)	1.28(3)	1.36(3)	1.13(1)
$\frac{\delta \langle m \rangle\|_{\rm cl}}{\delta \langle m \rangle\|_{\rm Q}}$	**None**	1.32(2)	1.31(2)	1.25(2)	1.10(2)	1.03(1)
	0.05	0.80(1)	0.80(1)	0.81(1)	0.96(1)	1.01(1)
	0.04	0.62(3)	0.69(3)	0.71(2)	0.93(1)	1.00(1)
	0.03	0.56(4)	0.57(5)	0.60(4)	0.87(1)	0.98(1)
	0.02	0.48(7)	0.49(7)	0.50(7)	0.77(2)	0.95(1)

TABLE I. The improvement found in various measurements performed using the Q-jet procedure compared to the classical pruning result, for a range of values of the rigidity parameter (α) and subject to a set of volatility cuts ($\mathcal{V} \leq \mathcal{V}_{\rm cut}$). The first set of rows exhibit the discovery potential $\langle S \rangle / \delta B$, while the second shows the average jet mass fluctuation $\delta \langle m \rangle$. In both cases results greater than unity indicate improvement over the classical pruning procedure (see the text for further discussion). For all quantities, the approximate statistical uncertainty for the last digit is shown in parenthesis.

With Q-jets, we can do something more sophisticated. Instead of the contribution of a given jet to S or B being 1 or 0 depending on whether the pruned mass is in the signal window or not, the contribution of the jet is now a rational number between 0 and 1, given by the fraction of the $N_{\rm tree}$ pruned masses that fall in the signal mass window. This is a way of reducing the contribution from events which are less signal like, without discarding them completely. In the limit $\alpha \to \infty$, this reduces to the classical measure, but for finite α, we expect an improvement in both significance and in $\delta \langle m \rangle$.

For numerical analysis we use the C/A algorithm for both the classical and Q-jets cases and take $N_{\rm tree} = 256$. (We find that the results saturate for $N_{\rm tree} \gtrsim 50$). We present results in Table I as ratios of the Q-jets result to the classical result, indicating the improvement in significance and mass uncertainty we can expect. These ratios should be independent of $\langle N_J \rangle$ and so we determine statistical uncertainties by fitting to results for $\langle N_J \rangle = 5, 10, 15$ and 20. The approximate statistical uncertainties are shown in parenthesis and apply to the last digit. We performed 10^4 pseudo-experiments, expecting $\mathcal{O}(1\%)$ statistical fluctuations from this procedure.

The first set of rows in Table I display measurements of the discovery potential $\langle S \rangle / \delta B$ compared to the re-

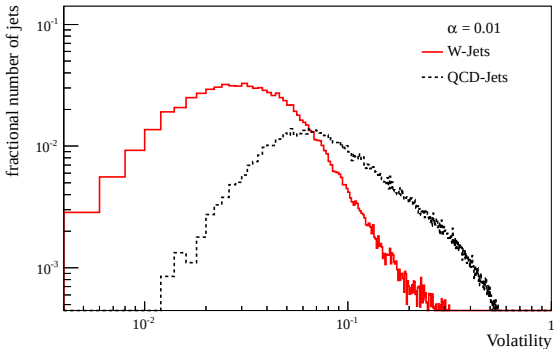

FIG. 2. Distribution of volatility for signal (boosted W-jets) and background (QCD jets) using a rigidity $\alpha = 0.01$.

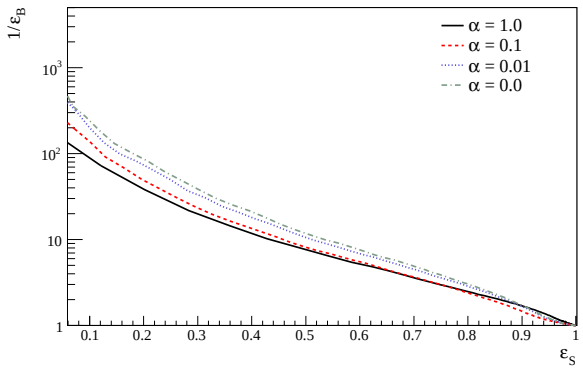

FIG. 3. The background versus signal efficiencies corresponding to a cut on volatility, for various α's, as compared to the classical pruning result.

sults with classical pruning. Focus on the rows labeled "none" for now (volatility is explained below). Since this quantity scales as $\sqrt{\mathcal{L}}$, the square of the number in the Table can be interpreted as an effective luminosity improvement due to employing the Q-jet procedure. For example, for $\alpha = 0.1$ the number 1.18 means an effective increase in the luminosity by $(1.18)^2 - 1 = 39\%$. Larger α values confine the range of trees and yield results very near the classical pruning results, i.e., $\frac{\langle S\rangle}{\delta B}\big|_{\rm Q} \to \frac{\langle S\rangle}{\delta B}\big|_{\rm cl}$. Smaller α values ($\alpha < 0.1$, with a much broader range of trees) also tend to degrade (decrease) the discovery potential.

The second set of rows exhibit the average jet mass fluctuation $\frac{\delta\langle m\rangle|_{\rm cl}}{\delta\langle m\rangle|_{\rm Q}}$ (note classical over Q-jets here). Values greater than unity mean that the mass can be measured more precisely with the Q-jet procedure for the same luminosity. Note that there is continuing improvement in $\delta\langle m\rangle$ as α decreases. That we get sensible results for (i.e.with a flat distance measure) is presumably because pruning is relatively insensitive to which tree we assign; even for physically unlikely clusterings, the hard radiation that reconstructs the mass is typically not pruned away.

The second way we have considered using Q-jets is in constructing qualitatively new types of observables. As an example, consider the **volatility** of a jet, defined by

$$\mathcal{V} = \Gamma/\langle m\rangle \ , \tag{4}$$

where $\Gamma \equiv \sqrt{\langle m^2\rangle - \langle m\rangle^2}$ and $\langle m\rangle$ are the RMS deviation and the mean of the pruned jet mass distribution for a single jet. The distribution of volatility for signal and background Q-jets with $\alpha = 0.01$ is shown in Fig. 2. We see that W jets have a lower volatility than QCD jets. This is easily understood, since the W jets have an intrinsic physical mass scale, while the QCD jets do not. Cutting on volatility, $\mathcal{V} \le \mathcal{V}_{\rm cut}$ can therefore improve significance in a boosted W search. The improvement is given in Table 1 for various values of $\mathcal{V}_{\rm cut}$.

The efficiencies for a volatility cut on signal and background are shown in Fig. 3. These efficiencies are defined as the fraction of the Q-jets that yield a pruned mass in the mass bin after the volatility cut. We plot them normalized to the classical results ($\alpha = \infty$ with no volatility cut). In the limit $\alpha \to \infty$ the curve collapses to the point (1,1). The upper right region of the plot corresponds to large values of $\mathcal{V}_{\rm cut}$, i.e., effectively no volatility cut. We find that the largest signal significance is obtained for a volatility cut of approximately 0.03, where for α near zero we achieve a relative $\langle S\rangle/\langle B\rangle$ of ~ 9 and a relative $\langle S\rangle/\delta B$ improvement of ~ 1.5 (the square of this number is the factor of two quoted in the Abstract). This corresponds to the neighborhood of the point $(0.25, 0.03)$ in Fig. 3. Finally we note that the precision of the mass measurement, shown in the lower rows in the table, is somewhat degraded by placing a cut on the volatility. This should not be a surprise as the cut discards some of the signal jets. A more comprehensive discussion of the statistics and of volatility will be given in [14].

In this paper, we have shown that it can be advantageous to consider a large number of trees constructed from the same jet in a single event, rather than a single tree as is done in traditional clustering algorithms. Although this paper has focused on tree-based observables, the Q-jets idea, of using non-determinism in event analysis, can naturally be applied in many other ways. Indeed, most observables, including jet substructure observables, such as jet masses, moments, pull [15], jet shapes [16], *etc.*, as well as more global observables, such as the number, distribution and 4-momenta for the jets in an event, work by trying to make the best guess at which properties of which final state particles tell us the most information about the underlying physics. The basic idea for Q-jets is that there is an inherent ambiguity in this best guess, both due to there not being a precise correspondence between final state particles and underlying physics, and due to our poor ability to extract that correspondence even if it were well-defined (as in a color singlet decay, for example). Thus, it would be natural to consider multiple interpretations of any observable, to

see whether getting away from the best guess can give us more robust information about the underlying physics, as it has with the tree-based substructure considered here. In will be interesting to see in future work how far this non-deterministic approach can be pushed.

SDE, AH, and TSR were supported in part by US Department of Energy under contract number DE-FGO2-96ER40956. MDS was supported in part by the Department of Energy, under grant DE-SC003916. DK was supported in part by a Simons postdoctoral fellowship and by an LHC-TI travel grant. AH, DK, and TSR were supported in part by the KITP, where a portion of this work was completed, under National Science Foundation under Grant No. PHY05-51164. Some computations were performed on the Odyssey cluster at Harvard University.

* sdellis@u.washington.edu
† ahornig@u.washington.edu
‡ tuhin@u.washington.edu
§ dkrohn@physics.harvard.edu
¶ schwartz@physics.harvard.edu

[1] Abdesselam, A. *et. al*, Eur.Phys.J., **C71**, 1661 (2011), arXiv:1012.5412 [hep-ph]; L. G. Almeida, R. Alon, and M. Spannowsky, (2011), arXiv:1110.3684 [hep-ph]; G. P. Salam, *ibid.*, **C67**, 637 (2010), arXiv:0906.1833 [hep-ph]; Altheimer, A. *et. al*, ArXiv e-prints (2012), arXiv:1201.0008 [hep-ph].

[2] S. D. Ellis and D. E. Soper, Phys.Rev., **D48**, 3160 (1993), arXiv:hep-ph/9305266 [hep-ph]; S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, Nucl. Phys., **B406**, 187 (1993).

[3] M. Wobisch and T. Wengler, (1998), arXiv:hep-ph/9907280 [hep-ph]; Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, JHEP, **08**, 001 (1997), arXiv:hep-ph/9707323.

[4] W. Giele and E. Glover, (1997), arXiv:hep-ph/9712355 [hep-ph].

[5] J. Gallicchio, J. Huth, M. Kagan, M. D. Schwartz, K. Black, *et al.*, JHEP, **1104**, 069 (2011), arXiv:1010.3698 [hep-ph]; Y. Cui, Z. Han, and M. D. Schwartz, Phys.Rev., **D83**, 074023 (2011), arXiv:1012.2077 [hep-ph]; D. E. Soper and M. Spannowsky, JHEP, **1008**, 029 (2010), arXiv:1005.0417 [hep-ph].

[6] D. E. Soper and M. Spannowsky, (2011), arXiv:1102.3480 [hep-ph].

[7] I. Volobouev, J.Phys.Conf.Ser., **293**, 012028 (2011).

[8] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, Phys.Rev., **D81**, 094023 (2010), arXiv:0912.0033 [hep-ph]; **D80**, 051501 (2009), arXiv:0903.5081 [hep-ph].

[9] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Phys.Rev.Lett., **100**, 242001 (2008), arXiv:0802.2470 [hep-ph]; D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, **101**, 142001 (2008), arXiv:0806.0848 [hep-ph]; D. Krohn, J. Thaler, and L.-T. Wang, JHEP, **1002**, 084 (2010), arXiv:0912.1342 [hep-ph].

[10] T. Sjostrand, S. Mrenna, and P. Z. Skands, JHEP, **0605**, 026 (2006), arXiv:hep-ph/0603175 [hep-ph].

[11] P. Z. Skands, Phys.Rev., **D82**, 074018 (2010), arXiv:1005.3457 [hep-ph].

[12] M. Cacciari, G. P. Salam, and G. Soyez, JHEP, **0804**, 063 (2008), arXiv:0802.1189 [hep-ph].

[13] M. Cacciari, G. Salam, and G. Soyez, "FastJet," Http://fastjet.fr/; M. Cacciari and G. P. Salam, Phys.Lett., **B641**, 57 (2006), arXiv:hep-ph/0512210 [hep-ph]; M. Cacciari, G. P. Salam, and G. Soyez, (2011), arXiv:1111.6097 [hep-ph].

[14] S. D. Ellis, A. Hornig, D. Krohn, T. S. Roy, and M. D. Schwartz, *in preparation*.

[15] J. Gallicchio and M. D. Schwartz, Phys. Rev. Lett., **105**, 022001 (2010), arXiv:1001.5027 [hep-ph].

[16] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, JHEP, **1011**, 101 (2010), arXiv:1001.0014 [hep-ph]; J. Gallicchio and M. D. Schwartz, Phys.Rev.Lett., **107**, 172001 (2011), arXiv:1106.3076 [hep-ph].