

Summer School on Symmetries, Fundamental Interactions and Cosmology September 2018, Fraueninsel/Chiemsee

Matteo Cacciari LPTHE Paris and Université Paris Diderot

lets

# Lecture 2 - Jet substructure

[Includes material from Gavin Salam and Grégory Soyez]





### Outline

## Jet algorithms

How are jets made

### Jet substructure

What's inside them

### What is a jet?



No, not this....

### A jet is something that happens in high energy events: a collimated bunch of hadrons flying roughly in the same direction

## Gluon 'discovery'



#### 1979:

**Three-jet events** observed by TASSO, JADE, MARK J and PLUTO at PETRA in e<sup>+</sup>e<sup>-</sup> collisions at 27.4 GeV

Interpretation: large angle emission of a hard gluon

## Jets viewed as a proxy to the initial partons



## Why jets

#### From PETRA to LEP

We could eyeball the collimated bunches, but it becomes impractical with millions of events

#### The classification of particles into jets is best done using a **clustering algorithm**



A few decades after PETRA and LEP, the event displays got prettier, but jets are still pretty much the same



#### Dijet event from CMS

## Jets @ LHC



#### 8(!) jets event from ATLAS

# Jets @ LHC



#### CMS Experiment at the LHC, CERN

Data recorded: 2011-Sep-12 14:31:09.077784 GMT(16:31:09 CEST) Run / Event: 176169 / 163029870

()-CERN.All right minimal

Marking and Appl

## Why do jets happen?



Gluon emission

$$\int \alpha_s \frac{dE}{E} \frac{d\theta}{\theta} \gg 1$$

Non-perturbative physics

 $\alpha_s \sim 1$ 

## The pervasiveness of jets

- ATLAS and CMS have each published 400+ papers since 2010
  - More than half of these papers make use of jets
  - 60% of the searches papers makes use of jets



(Source: INSPIRE. Results may vary when employing different search keywords)

## Taming reality



One purpose of a 'jet clustering' algorithm is to reduce the complexity of the final state, simplifying many hadrons to simpler objects that one can hope to calculate

## Jet definitions as projections



Projection to jets should be resilient to QCD effects

#### NB: projections are NOT unique: a jet is NOT EQUIVALENT to a parton



#### 2 clear jets

3 jets?



#### 2 clear jets

#### 3 jets? or 4 jets?

Gavin Salam (CERN)

QCD basics 4

## Jet clustering algorithm

A **jet algorithm** maps the momenta of the final state particles into the momenta of a certain number of jets:



4-momenta, calorimeter towers, ....

# Most algorithms contain a resolution parameter, **R**, which controls the extension of the jet

"Jet [definitions] are legal contracts between theorists and experimentalists" -- MJ Tannenbaum

### Jets can serve two purposes

- They can be observables, that one can measure and calculate
- They can be tools, that one can employ to extract specific properties of the final state

Different clustering algorithms have different properties and characteristics that can make them more or less appropriate for each of these tasks

## **IRC** safety

An observable is **infrared and collinear safe** if, in the limit of a **collinear splitting**, or the **emission of an infinitely soft** particle, the observable remains **unchanged**:

$$O(X; p_1, \dots, p_n, p_{n+1} \to 0) \to O(X; p_1, \dots, p_n)$$
  
$$O(X; p_1, \dots, p_n \parallel p_{n+1}) \to O(X; p_1, \dots, p_n + p_{n+1})$$

This property ensures cancellation of **real** and **virtual** divergences in higher order calculations

If we wish to be able to calculate a jet rate in perturbative QCD the jet algorithm that we use must be IRC safe: soft emissions and collinear splittings must not change the hard jets

## Reconstructing jets must respect rules



#### Perturbative calculations of jet observable will only be possible with collinear (and infrared) safe jet definitions

## **Cone algorithms**

The first rigorous definition of cone jets in QCD is due to Sterman and Weinberg Phys. Rev. Lett. **39**, 1436 (1977)





3 particles 2 ) collinear Soft 2 jets Parge angle 3 jets

## Jet algorithms

The Sterman-Weinberg definition is "inclusive enough" for IRC safety Good for 2 jets and e<sup>+</sup>e<sup>-</sup> collisions

What happens in a more general case, where more than two jets are likely to exist? **Where** do we place the cones? **How many**?

#### **Iterative** jet algorithms

#### In UED these are usually call

#### In HEP these are usually called **cone** and **sequential recombination** algorithms respectively

(in other fields they are often called partitional-type clustering and agglomerative hierarchical clustering)

### Two main approaches to jet clustering

- I. Find regions where a lot of energy flows
- 2. Decide which particles are "close", aggregate them



## Two main classes of jet algorithms

#### Sequential recombination algorithms

Bottom-up approach: combine particles starting from **closest ones** How? Choose a **distance measure**, iterate recombination until few objects left, call them jets

> Works because of mapping closeness ⇔ QCD divergence Examples: Jade, kt, Cambridge/Aachen, anti-kt, .....

> > Usually trivially made IRC safe, but their algorithmic complexity scales like N<sup>3</sup>

#### Cone algorithms

Top-down approach: find coarse regions of energy flow.

**How?** Find **stable cones** (i.e. their axis coincides with sum of momenta of particles in it) Works because QCD only modifies energy flow on small scales Examples: JetClu, MidPoint, ATLAS cone, CMS cone, SISCone.....

Can be programmed to be fairly fast, at the price of being complex and IRC unsafe

### Snowmass

#### FERMILAB-Conf-90/249-E [E-741/CDF]

#### Toward a Standardization of Jet Definitions ·

\* To be published in the proceedings of the 1990 Summer Study on High Energy Physics, *Research Directions for the Decade*, Snowmass, Colorado, June 25 - July 13, 1990.



## A little history

- Cone-type jets were introduced first in QCD in the 1970s (Sterman-Weinberg '77)
- In the 1980s cone-type jets were adapted for use in hadron colliders (SppS, Tevatron...) → iterative cone algorithms
- LEP was a golden era for jets: new algorithms and many relevant calculations during the 1990s
  - Introduction of the 'theory-friendly' kt algorithm
    - sequential recombination type algorithm, IRC safe
    - it allows for all order resummation of jet rates
  - Several accurate calculations in perturbative QCD of jet properties: rates, jet mass, thrust, ....

## Finding stable cones

#### In partitional-type algorithms (i.e. cones), one wishes to find the **stable configurations**:

axis of cones coincides with sum of 4-momenta of the particles it contains

#### The 'safe' way of doing so is to test **all possible combinations** of N objects

Unfortunately, this takes N2<sup>N</sup> operations: the time taken is the age of the universe for only 100 objects

An approximate way out is to use **seeds** (e.g. à la k-means) However, the final result can depend on the choice of the seeds and, such jet algorithms usually turn out to be **IRC unsafe** 

## Finding cones

Different procedures for placing the cones lead to **different cone algorithms** 

NB: their properties and behaviour can **vastly differ**: there isn't **'a'** cone algorithm, but rather many of them

The main sub-categories of cone algorithms are:

Fixed cone with progressive removal (FC-PR) (PyJet, CellJet, GetJet)
Iterative cone with progressive removal (IC-PR) (CMS iterative cone)
Iterative cone with split-merge (IC-SM) (JetClu, ATLAS cone)
IC-SM with mid-points (ICmp-SM) (CDF MidPoint, D0 Run II)
ICmp with split-drop (ICmp-SD) (PxCone)
Seedless cone with split-merge (SC-SM) (SISCone)

#### All, except SISCone, are infrared or collinear unsafe

## **Recombination algorithms**

- First introduced in e<sup>+</sup>e<sup>-</sup> collisions in the '80s
- Typically they work by calculating a 'distance' between particles, and then recombine them pairwise according to a given order, until some condition is met (e.g. no particles are left, or the distance crosses a given threshold)

#### IRC safety can usually be seen to be trivially guaranteed

## JADE algorithm

Distance: 
$$y_{ij} = rac{2E_iE_j(1-\cos heta_{ij})}{Q^2}$$

- Find the minimum y<sub>min</sub> of all y<sub>ij</sub>
- If y<sub>min</sub> is below some jet resolution threshold y<sub>cut</sub>, recombine i and j into a single new particle ('pseudojet'), and repeat
- If no  $y_{min} < y_{cut}$  are left, all remaining particles are jets

Problem of this particular algorithm: two **soft** particles emitted at **large angle** get easily recombined into a single jet: counterintuitive and perturbatively troublesome

## e<sup>+</sup>e<sup>-</sup> k<sub>t</sub> (Durham) algorithm

[Catani, Dokshitzer, Olsson, Turnock, Webber '91]

Distance: 
$$y_{ij} = \frac{2\min(E_i^2, E_j^2)(1 - \cos\theta_{ij})}{Q^2}$$

In the collinear limit, the numerator reduces to the **relative transverse momentum** (squared) of the two particles, hence the name of the algorithm

- Find the minimum y<sub>min</sub> of all y<sub>ij</sub>
- If y<sub>min</sub> is below some jet resolution threshold y<sub>cut</sub>, recombine i and j into a single new particle ('pseudojet'), and repeat
- If no  $y_{min} < y_{cut}$  are left, all remaining particles are jets

## e<sup>+</sup>e<sup>-</sup> k<sub>t</sub> (Durham) algorithm in action



Resummed calculations for distributions of  $y_{cut}$  doable with the  $k_t$  algorithm

## e<sup>+</sup>e<sup>-</sup> k<sub>t</sub> (Durham) algorithm v. QCD

#### kt is a sequential recombination type algorithm

One key feature of the k<sub>t</sub> algorithm is its relation to the structure of QCD divergences:

$$\frac{dP_{k\to ij}}{dE_i d\theta_{ij}} \sim \frac{\alpha_s}{\min(E_i, E_j)\theta_{ij}}$$

The  $y_{ij}$  distance is the inverse of the emission probability

The kt algorithm roughly inverts the QCD branching sequence (the pair which is recombined first is the one with the largest probability to have branched)

The history of successive clusterings has physical meaning

## The common wisdom circa 2005

#### Cone algorithms are IRC unsafe

- ➡ because, to make them reasonably fast, they were usually implemented via approximate methods using seeds
- Sequential recombination algorithms (i.e. k<sub>t</sub>) are slow and too susceptible to background contamination
  - $\rightarrow$  because they scale like N<sup>3</sup>
  - because they tend to collect soft particles up to large distances from centre
  - → because they were often run with R=1 and compared to cones with R=0.5!

## Geometry

The solution to the speed problem came from considering the clustering problem from a geometrical rather from a combinatorial point of view

Sequential recombination algorithms could be implemented with O(N<sup>2</sup>) or even O(NInN) complexity rather than O(N<sup>3</sup>) [MC, Salam, 2006]

Cone algorithms could be implemented exactly (and therefore made IRC safe) with O(N<sup>2</sup>InN) rather than O(N2<sup>N</sup>) complexity [Salam, Soyez 2007]

## kt algorithm in hadron collisions

(Inclusive and longitudinally invariant version)

$$d_{ij} = \min(p_{ti}^2, p_{tj}^2) \frac{\Delta R_{ij}^2}{R^2} \qquad \qquad d_{iB} = p_{ti}^2$$

- Calculate the distances between the particles: **d**<sub>ij</sub>
- Calculate the beam distances: **d**<sub>iB</sub>
- Combine particles with smallest distance d<sub>ij</sub> or, if d<sub>iB</sub> is smallest, call it a jet
- Find again smallest distance and repeat procedure until no particles are left (this stopping criterion leads to the *inclusive* version of the k<sub>t</sub> algorithm)
- Only use jets with pt > pt,min

## The speed 'problem'

Given N particles the k<sub>t</sub> algorithm is, naively, an **O(N<sup>3</sup>) algorithm**: calculate N<sup>2</sup> distances, repeat for all N iterations

# With 1000 particles, this takes 10<sup>9</sup> operations, i.e. **about a second**

Clustering such an event would take significantly more than generating it in a MonteCarlo, not to speak about trying to use it at the trigger level, where the time budges is of the order of **tens of milliseconds** 

This, together with the tendency of the k<sub>t</sub> algorithm to 'scoop up' soft radiation quite far from the hard partons, and to give jets with ragged borders, difficult to correct for, had led people to prefer cones in a hadronic environment

## The FastJet lemma

To improve the speed of the algorithm we must determine more efficiently which particle is "close" to another and therefore gets combined with it

#### Key lemma:

[MC, G.P. Salam, hep-ph/0512210]

If i and j form the **smallest**  $d_{ij} = min(p_{ti}^2, p_{tj}^2) \Delta R_{ij}^2$  and  $p_{ti} < p_{tj}$ 

 $\Rightarrow \quad \Delta R_{ij} < \Delta R_{ik} \quad \forall k \neq j$ 

i.e. j is the **geometrical** nearest neighbour of i

#### Proof:

Suppose the lemma wrong:  $\exists k \text{ such that } \Delta R_{ij} \geq \Delta R_{ik}$ 

then  $d_{ik} = \min(p_{ti}^2, p_{tk}^2) \Delta R_{ik}^2 \le p_{ti}^2 \Delta R_{ik}^2 \le p_{ti}^2 \Delta R_{ij}^2 = d_{ij}$ 

which contradicts the initial statement that  $d_{ij}$  is the smallest one

## The FastJet lemma

Translation from mathematics:

When a particle gets combined with another, and has the smallest  $k_t$ , its partner will be its **geometrical nearest neighbour** on the cylinder spanned by y and  $\varphi$ 

This means that we need to look for partners only among the O(N) nearest neighbours of all particles (a few neighbours each × N particles)

## The FastJet algorithm

- For each particle i establish its geometrical nearest neighbour G<sub>i</sub> and calculate the arrays d<sub>iGi</sub> and d<sub>iB</sub>
- Find the minimal value d<sub>min</sub> of the d<sub>iGi</sub> and d<sub>iB</sub>, combine particles corresponding to it
- ▶ Update **d**<sub>iGi</sub> and **d**<sub>iB</sub> if needed, continue until no particles left

#### This is already an $O(N^2)$ algorithm: i.e. find O(N) neighbours for N particles

#### But we can do better

## The FastJet algorithm

Our problem has now become a **geometrical** one: how to find efficiently the (nearest) neighbour(s) of a point

Widely studied problem in computational geometry Tool: **Voronoi diagram** 

Definition: each cell contains the locations which have the given point as nearest neighbour



The dual of a Voronoi diagram is a **Delaunay triangulation** 

Key feature: once the Voronoi diagram is constructed, <u>the nearest neighbour of a</u> point will be in one of the O(1) cells sharing an edge with its own cell

Example : the G(eometrical) N(earest) N(eighbour) of point 7 will be found among 1,4,2,8 and 3 (it turns out to be 3)

### The FastJet algorithm

MC and G.P. Salam, hep-ph/0512210

O(N InN)

O(N InN)

Construct the Voronoi diagram of the N particles using the CGAL library

Find the GNN of each of the N particles. Construct the  $d_{ij}$  distances, store the results in a priority queue (C++ map)

Merge/eliminate particles appropriately

Update Voronoi diagram and distances' map O(InN)

### Overall, an O(N In N) algorithm

PRISMA Summer School - September 2018



## The kt algorithm and its siblings

$$d_{ij} = \min(p_{ti}^{2p}, p_{tj}^{2p}) \frac{\Delta y^2 + \Delta \phi^2}{R^2} \qquad d_{iB} = p_{ti}^{2p}$$

**P** = k<sub>t</sub> algorithm S. Catani, Y. Dokshitzer, M. Seymour and B. Webber, Nucl. Phys. B406 (1993) 187 S.D. Ellis and D.E. Soper, Phys. Rev. D48 (1993) 3160

#### **p=0** Cambridge/Aachen algorithm Y. Dokshitzer, G. Leder, S.Moretti and B. Webber, JHEP 08 (1997) 001 M.Wobisch and T.Wengler, hep-ph/9907280

#### **p** = - I anti-k<sub>t</sub> algorithm

MC, G. Salam and G. Soyez, arXiv:0802.1189

NB: in anti-kt pairs with a **hard** particle will cluster first: if no other hard particles are close by, the algorithm will give **perfect cones** 

Quite ironically, a sequential recombination algorithm is the 'perfect' cone algorithm

## IRC safety of generalised-kt algorithms

$$d_{ij} = \min(p_{ti}^{2p}, p_{tj}^{2p}) \frac{\Delta y^2 + \Delta \phi^2}{R^2} \qquad d_{iB} = p_{ti}^{2p}$$

#### p > 0

New **soft** particle  $(p_t \rightarrow 0)$  means that  $d \rightarrow 0 \Rightarrow$  clustered first, no effect on jets New **collinear** particle  $(\Delta y^2 + \Delta \Phi^2 \rightarrow 0)$  means that  $d \rightarrow 0 \Rightarrow$  clustered first, no effect on jets

#### **p** = 0

New **soft** particle  $(p_t \rightarrow 0)$  can be new jet of zero momentum  $\Rightarrow$  no effect on hard jets New **collinear** particle  $(\Delta y^2 + \Delta \Phi^2 \rightarrow 0)$  means that  $d \rightarrow 0 \Rightarrow$  clustered first, no effect on jets

#### p < 0

New **soft** particle  $(p_t \rightarrow 0)$  means  $d \rightarrow \infty \Rightarrow$  clustered last or new zero-jet, no effect on hard jets New **collinear** particle  $(\Delta y^2 + \Delta \Phi^2 \rightarrow 0)$  means that  $d \rightarrow 0 \Rightarrow$  clustered first, no effect on jets

	IRC safe algorithms		
kt	$SR d_{ij} = min(p_{ti}^2, p_{tj}^2) \Delta R_{ij}^2 / R^2 hierarchical in rel p_t$	Catani et al '91 Ellis, Soper '93	NInN
Cambridge/ Aachen	$SR \\ d_{ij} = \Delta R_{ij}^2 / R^2 \\ hierarchical in angle$	Dokshitzer et al '97 Wengler, Wobish '98	NInN
anti-k <sub>t</sub>	SR d <sub>ij</sub> = min(p <sub>ti</sub> - <sup>2</sup> ,p <sub>tj</sub> - <sup>2</sup> )ΔR <sub>ij</sub> <sup>2</sup> /R <sup>2</sup> gives perfectly conical hard jets	MC, Salam, Soyez '08 (Delsart, Loch)	N <sup>3/2</sup>
SISCone	Seedless iterative cone with split-merge gives 'economical' jets	Salam, Soyez '07	N²InN
<b>'second-generation' algorithms</b> All are available in FastJet, <u>http://fastjet.fr</u> (As well as many IRC unsafe ones)			

Matteo Cacciari - LPTHE

## FastJet speed

#### Time needed to cluster an event with N particles



Matteo Cacciari - LPTHE

PRISMA Summer School - September 2018







Clustering grows around hard cores  $d_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$ 



Clustering grows around hard cores  $d_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$ 











Clustering grows around hard cores  $d_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$ 



Clustering grows around hard cores  $d_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$ 



Clustering grows around hard cores  $d_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$ 



Clustering grows around hard cores  $d_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$ 



Clustering grows around hard cores  $d_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$ 



Clustering grows around hard cores

$$_{ij} = \frac{1}{\max(p_{ti}^2, p_{tj}^2)} \frac{\Delta R_{ij}^2}{R^2}, \quad d_{iB} = \frac{1}{p_{ti}^2}$$



d

Anti-kt gives circular jets ("cone-like") in a way that's infrared safe









### Example of jet observable



## Recap of Lecture I

- A vast zoology of jet algorithms has been reduced in the past few years to 4 infrared and collinear safe algorithms
  - All are implemented in an efficient and fast way
  - Of these, anti-kt is used by all the LHC collaborations as their main algorithm for "finding" jets and measuring inclusive cross sections
- The four algorithms have quite different characteristics, which makes them non easily swappable when specific properties are needed for specific tasks. On the other hand, chances are that one can chose the algorithm which is most appropriate for a specific job